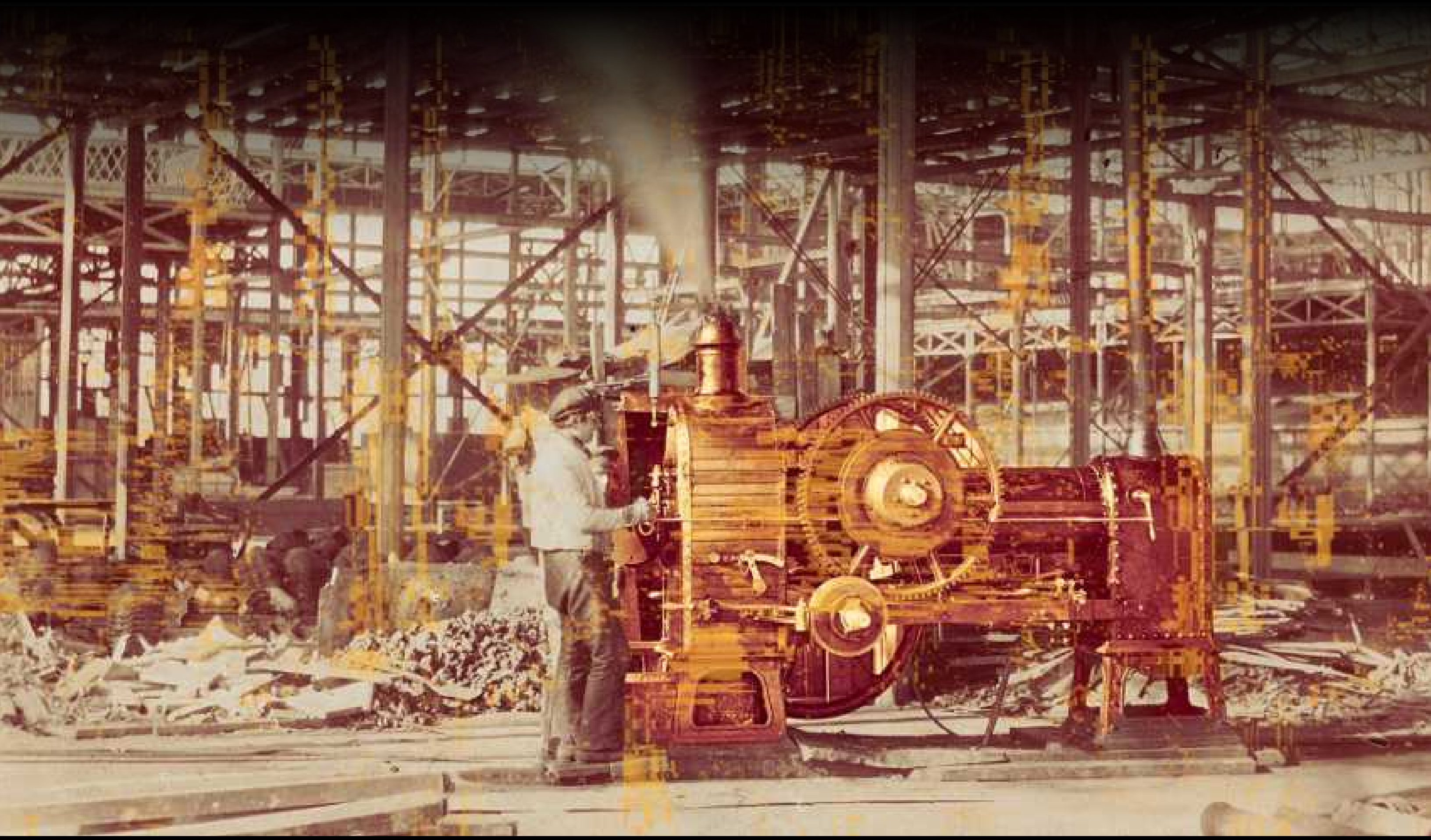
UC BERKELEY

CENTER FOR LONG-TERM CYBERSECURITY



General-Purpose AI Risk-Management Standards Profile

NADA MADKOUR | JESSICA NEWMAN | DEEPIKA RAMAN | KRYSTAL JACKSON
EVAN R. MURPHY | CHARLOTTE YUAN | DAN HENDRYCKS

Version 1.2, April 2026

For the latest public version of this document, see: https://cltc.berkeley.edu/publication/ai-risk-management-standards-profile-v1.2

For the Agentic AI Risk-Management Standards Profile , see: https://cltc.berkeley.edu/publication/agentic-ai-risk-profile/

For the QUICK GUIDE: An Introductory Resource for the General-Purpose AI Risk-Management Standards Profile V1.2, see: https://cltc.berkeley.edu/wp-content/uploads/2026/04/Berkeley-Profile-v1-2-Quick-Guide.pdf

For Evaluation of Frontier AI Company Practices Using the General-Purpose AI Risk-Management Standards Profile V1.2, see https://cltc.berkeley.edu/wp-content/uploads/2026/04/Berkeley-Evaluation-of-Frontier-AI-v1-2.pdf

For Mapping Key Standards and Regulations to the General-Purpose AI Risk Management Standards Profile V1.2, see: https://cltc.berkeley.edu/wp-content/uploads/2026/04/Berkeley-Mapping-of-Profile-Guidance-v1-2-to-Key-Standards-and-Regulations.pdf

For Transparency, Documentation, and Reporting Recommendations for General-Purpose AI Risk-Management, see: https://cltc.berkeley.edu/wp-content/uploads/2026/04/Berkeley-Transparency-Recommendations-for-GPAI-Risk-Management.pdf

Cover art: The cover image is an adaptation of a photograph titled, "Steam Engine near the Grand Transept, Crystal Palace," taken by the photographer Philip Henry Delamotte in 1851. The impact of artificial intelligence and especially general purpose artificial intelligence is often compared to the impact of the steam engine during the Industrial Revolution, which brought enormous economic gains, but also dangerous workplaces and horrible living conditions for many. The Crystal Palace housed the Great Exhibition of 1851, where examples of technology developed in the Industrial Revolution were put on display for thousands of people to see. While enjoyed by many, the Crystal Palace was also critiqued for representing a false utopia. Similarly, the rise of general purpose AI is often discussed with utopian visions, but such positive visions are often overpromised and will not be possible without the establishment of meaningful risk management strategies. The image is a reminder of the entanglement of people and machines, and the profound and lasting impact of general purpose technologies on society.

In this adaptation, the updated golden palette alludes to contemporary narratives of an "AI gold rush," reflecting the rapid investment, aspiration, and momentum surrounding AI development. The radiant gold machinery draws the viewer's eye and underscores how technological systems increasingly occupy the locus of attention within public and policy discourse, often overshadowing the human figure within the frame. Against this backdrop of acceleration and possibility, we present the second annual update to the AI Risk-Management Standards Profile for General-Purpose AI Systems (GPAIS) and Foundation Models (Version 1.2).

General-Purpose AI Risk-Management Standards Profile

NADA MADKOUR† • JESSICA NEWMAN† • DEEPIKA RAMAN† • KRYSTAL JACKSON†
EVAN R. MURPHY† • CHARLOTTE YUAN† • DAN HENDRYCKS††

† AI Security Initiative, Center for Long-Term Cybersecurity, UC Berkeley
†† Berkeley AI Research Lab, UC Berkeley
All affiliations listed are either current, or were during main contributions to this work or a previous version.

Version 1.2, April 2026

ABSTRACT

General-purpose AI (GPAI) models, including cutting-edge large language models and agentic models, can provide many beneficial capabilities, but also introduce risks of adverse events with profound consequences. This document provides risk-management practices or controls for identifying, analyzing, and mitigating risks of GPAI models. We intend this document primarily for developers of large-scale, state-of-the-art GPAI models, including for research projects or internally deployed models that are not made available to users outside of an organization. Others who can benefit from this guidance include downstream developers and deployers of end-use applications that build on a GPAI model. This document facilitates conformity with or use of leading standards related to AI risk-management, adapting and building on the voluntary guidance in the NIST AI Risk-Management Framework and ISO/IEC 23894, with a focus on the unique issues faced by developers and deployers of GPAI models.

NOTES ON THIS VERSION

Changes between this Version 1.2 Profile and the Version 1.1 Profile (Barrett et al. 2025) include:

- The addition of two high-priority subcategories:
    - **Govern 5.1:** Recognizing the critical role of external feedback, especially third-party evaluations, in robust AI risk management, we included a dedicated sub-category to emphasize the importance of this essential step.
    - **Manage 4.1:** Risk management does not end following a model's deployment; continuous monitoring is required. Therefore, this update includes the addition of post-deployment monitoring as a high-priority sub-category.
- The addition of Section 2.2.1 on risk taxonomies.
- Terminology and scope refinements throughout this document:
    - Most notable is that most instances of "general-purpose AI (GPAI)/foundation models" were changed to "GPAI models" to simplify our terminology.
- Additional, or updated, resources for:
    - Establishing and operationalizing risk thresholds (Map 1.5);
    - Red-teaming and benchmark capability evaluations (Measure 1.1);
    - Transparency and documentation (Govern 1.4, Measure 2.9 and 3.1); and
    - Incident response plans (Govern 1.4).
- Added actions and guidance from the EU GPAI Code of Practice (EC 2025a) under several sub-categories.
- Updated resources to their latest versions — e.g., the International AI Safety Report (Bengio et al. 2025) and NIST AI 800-1 2pd (NIST 2025).
- Expansion of AI risks, including new content on:
    - Manipulation and deception (Map 5.1);
    - Sandbagging during evaluations of hazardous capabilities (Govern 2.1, Map 5.1);
    - Situational awareness (Map 5.1);
    - Socioeconomic and labor market disruption (Map 5.1); and
    - Possible intractability of removing backdoors (Map 5.1, Measure 2.7).
- The Roadmap of Issues to Address in Future Versions of the Profile (Appendix 3) has been updated to include:
    - Restructuring the GPAI Profile around a pre-defined risk taxonomy, such as the one proposed in section 2.2.1;
    - Developing targeted supplementary guidance for different stakeholders;
    - Expanded guidance on internal deployment and AI R&D; and
    - Enhanced risk-to-mitigation mapping.

- Updates to the **Mapping Key Standards and Regulations to the AI Risk-Management Standards Profile for General-Purpose AI V1.2 document**. To access, see: https://cltc.berkeley.edu/wp-content/uploads/2026/04/Berkeley-Mapping-of-Profile-Guidance-v1-2-to-Key-Standards-and-Regulations.pdf
  - Added mapping to the EU GPAI Code of Practice (EC 2025a).
- Updates to the **Evaluation of Frontier AI Company Practices Using the General-Purpose AI Risk-Management Standards Profile** (formerly Retrospective Test Use of Profile Guidance). To access, see: https://cltc.berkeley.edu/wp-content/uploads/2026/04/Berkeley-Evaluation-of-Frontier-AI-v1-2.pdf
  - Testing on new GPAI models (Claude Opus 4.5, GPT-5, Llama 4, and Gemini 3 Pro).
  - To facilitate comparison, we included V1.1 testing results alongside current (V1.2) testing results for each model (e.g., Claude, Llama, GPT, and Gemini).
- New supporting documentation:
  - **Agentic AI Risk-Management Standards Profile.** To access, please see: https://cltc.berkeley.edu/publication/agentic-ai-risk-profile/
  - **Transparency, Documentation, and Reporting Recommendations for General-Purpose AI Risk-Management.** To access, please see: https://cltc.berkeley.edu/wp-content/uploads/2026/04/Berkeley-Transparency-Recommendations-for-GPAI-Risk-Management.pdf

Contents

Note: Subcategories under Section 3 (Guidance) that are referenced as "high-priority risk-management steps" are included for ease of reference.

Executive Summary

Increasingly multi-purpose AI models, such as state-of-the-art large language models (LLMs) or other **general-purpose AI (GPAI) models**, **generative AI models, and frontier models** (typically all referred to hereafter with **the umbrella term "GPAI models"),** can provide many beneficial capabilities, but they also introduce risks of adverse events with harmful impacts at societal scale.

This document provides an AI risk-management standards **profile (hereafter "Profile")**, a targeted set of risk-management practices or controls specifically for identifying, analyzing, and mitigating risks of GPAI models. This Profile is designed to complement the broadly applicable guidance in the NIST AI risk-management Framework (AI RMF) or related AI risk-management standards, such as ISO/IEC 23894.

We intend for this Profile to be used primarily by **developers of large-scale, state-of-the-art GPAI models**. For GPAI model developers, this Profile facilitates conformity with or use of leading AI risk-management-related standards, and aims to facilitate compliance with relevant regulations such as the EU AI Act, especially for aspects related to GPAI models. However, this Profile does not provide all guidance that may be needed for applications or AI systems that incorporate GPAI models in particular industry sectors or applications. Additionally, given the nascent and rapidly evolving state of GPAI safety science and GPAI security, these practices should be understood as essential but not necessarily sufficient for achieving acceptable risk levels.

Others who can benefit from the use of this guidance include downstream developers of end-use applications or AI systems that build on a GPAI model; evaluators of GPAI models; and the regulatory community. This Profile can provide GPAI model deployers, evaluators, and regulators with information useful for evaluating the extent to which developers of GPAI models have followed relevant best practices. Normalizing the use of best practices such as those detailed in this Profile can help ensure developers of GPAI models can be competitive without compromising on practices for AI safety, security, accountability, and related issues.

Ultimately, this Profile aims to help key actors in the value chains of GPAI models and systems to maximize benefits, and minimize negative impacts, to individuals, communities, organizations, society, and the planet. That includes protection of human rights, minimization of negative environmental impacts, and prevention of adverse events with systemic or catastrophic consequences at societal scale.

The NIST AI RMF "core functions," or broad categories of activities, apply as appropriate across AI system lifecycles, and we provide corresponding guidance in related sections of this Profile: "Govern" (Section 3.1) for AI risk-management policies, roles, and responsibilities; "Map" (Section 3.2) for identifying AI risks in context; "Measure" (Section 3.3) for rating AI trustworthiness characteristics; and "Manage" (Section 3.4) for decisions on prioritizing, avoiding, mitigating, or accepting AI risks.

Users of this Profile should place high priority on the following risk-management steps and corresponding guidance sections. (Appropriately applying the guidance for the following steps should be regarded as the baseline or minimum expectations for users of this Profile, who can exceed the minimum expectations by also applying guidance in other sections.)

- **Check or update, and incorporate, each of the high-priority risk-management steps when making go/no-go decisions,** especially on whether to proceed on major stages or investments for development or deployment of cutting-edge large-scale GPAI models (Manage 1.1).
- **Take responsibility for risk assessment and risk-management tasks for which your organization has access to information, possesses requisite resources, or has the opportunity to develop capabilities sufficient for constructive action, especially when these are substantially greater than that of others in the value chain** (Govern 2.1).
- We also recommend applying this principle throughout other risk assessment and risk-management steps, and we refer to it frequently in other guidance sections.
- **Collect, consider, and integrate feedback from external stakeholders and third-party evaluators** (Govern 5.1).
    - Prioritize partnerships with independent third-party evaluators and auditors throughout the model lifecycle, particularly when assessing the model for dangerous capabilities, and establish clear and accessible feedback channels for users or impacted persons.
- **Set risk-tolerance thresholds to prevent unacceptable risks** (Map 1.5).
    - For example, the NIST AI RMF 1.0 recommends the following: "In cases where an AI system presents unacceptable negative risk levels — such as where significant negative impacts are imminent, severe harms are actually occurring, or catastrophic risks are present — development and deployment should cease in a safe manner until risks can be sufficiently managed" (NIST 2023a, p.8).

- **Identify reasonably foreseeable uses, misuses, and abuses for a GPAI model** (e.g., automated generation of toxic or illegal content or disinformation, or aiding with proliferation of cyber, chemical, biological, radiological, or nuclear weapons), and identify reasonably foreseeable potential impacts (e.g., to fundamental rights) (Map 1.1).
- **Identify whether a GPAI model could lead to significant, severe, or catastrophic impacts,** e.g., due to correlated failures or errors across high-stakes deployment domains, dangerous emergent behaviors or vulnerabilities, or harmful misuses and abuses (Map 5.1).
- **Use red teams and adversarial testing** as part of extensive interaction with GPAI models to identify dangerous capabilities, vulnerabilities, or other emergent properties of such systems (Measure 1.1).
- **Track important identified risks** (e.g., vulnerabilities from data poisoning and other attacks or mis-specification of objectives), even if they cannot yet be measured (Measure 1.1 and Measure 3.2).
- **Implement risk-reduction controls as appropriate** throughout the GPAI model lifecycle, e.g., through independent auditing, incremental scale-up, red-teaming, structured access or staged release, and other steps (Manage 1.3, Manage 2.3, and Manage 2.4).
- **Incorporate identified AI system risk factors — and circumstances that could result in impacts or harms — into reporting and engagement with internal and external stakeholders** (e.g., when reporting to downstream developers, regulators, users, impacted communities, etc.), using model cards, system cards, and other transparency mechanisms (Govern 4.2).
- **Continuously monitor the model post-deployment** by actively gathering relevant information for risk evaluation and implementing an incident response plan (Manage 4.1).

We also recommend **documenting the process used in considering risk mitigation controls, the options considered, and the reasons for choices made**. Documentation on many items should be shared in publicly available materials such as system cards. Details on particular issues, such as security vulnerabilities, can be responsibly omitted from public materials to reduce misuse potential, especially if available to auditors, Information Sharing and Analysis Organizations, or other parties.

GPAI model-related risk topics and corresponding guidance sections in this Profile include the following. (Some of these topics overlap, in part because the guidance often involves iterative assessments for additional depth on issues identified at earlier stages.)

- Reasonably foreseeable impacts (Section 3.2, Map 1.1), including impacts:
  - To individuals, including to health, safety, well-being, or fundamental rights;
  - To groups, including populations vulnerable to disproportionate adverse impacts or harms; and
  - To society, including environmental impacts.
- Significant, severe, or catastrophic harm factors (Section 3.2, Map 5.1), including:
  - Correlated bias and discrimination;
  - Impacts to societal trust or democratic processes;
  - Correlated robustness failures;
  - Potential for high-impact misuses, such as for cyber weapons, or chemical, biological, radiological, or nuclear (CBRN) weapons;
  - Capability to manipulate or deceive humans in harmful ways; and
  - Loss of understanding and control of an AI system in a real-world context.
- AI trustworthiness characteristics (Section 3.4, Measure 2), including:
  - Safety, reliability, and robustness (Measure 2.5, Measure 2.6);
  - Security and resiliency (Measure 2.7);
  - Accountability and transparency (Measure 2.8);
  - Explainability and interpretability (Measure 2.9);
  - Privacy (Measure 2.10); and
  - Fairness and bias (Measure 2.11).

Additional topics to address in future versions of the Profile are listed in Appendix 3.

1. Introduction and Objectives

1.1 KEY TERMS

Increasingly multi-purpose AI models, such as state-of-the-art large language models (LLMs), large multimodal language models (LMMs), or other general-purpose AI (GPAI) models, frontier models, and generative AI models, can provide many beneficial capabilities, but they also introduce risks of adverse events with potential consequences at societal scale.

Below is an overview of key terms used in this Profile. (For additional terms and acronyms, see the Glossary.)

- **General-Purpose AI (GPAI):** Our usage of the terms "general-purpose AI model" and "general-purpose AI system" is very similar to the corresponding terms in the EU AI Act (EP 2024), except that we do not exclude AI models used for research.
  - **GPAI Model:** *"General-purpose AI model' means an AI model, including where such an AI model is trained with a large amount of data using self-supervision at scale, that displays significant generality and is capable of competently performing a wide range of distinct tasks regardless of the way the model is placed on the market and that can be integrated into a variety of downstream systems or applications ...*" (EP 2024, Article 3(63)).
    - Examples of GPAI models include GPT-5, Claude 4, PaLM 2, LLaMA 3, and others.
  - **GPAI System:** "*General-purpose AI system' means an AI system which is based on a general-purpose AI model and which has the capability to serve a variety of purposes, both for direct use as well as for integration in other AI systems*" (EP 2024, Article 3(66)).

- **Frontier model**: A cutting-edge, state-of-the-art, or highly capable GPAI model. Such models also may possess hazardous or dual-use capabilities sufficient to pose severe risks to public safety. (See, e.g., Ganguli, Hernandez et al. 2022, Anderljung, Barnhart et al. 2023, and Microsoft 2023.)
  - We treat **frontier models as the largest-scale, highest-capability subset of GPAI models.** They are typically characterized by model size, training compute or data, and/or resulting capabilities that are above or near to industry-record thresholds. (See also "foundation model frontier" in the Glossary.)
  - Our usage of the term "frontier model" approximately corresponds to dual-use foundation models, as defined by former Executive Order 14110 (Federal Register 2023) and to GPAI models with systemic risk, as defined by the EU AI Act (EP 2024).

- Examples of frontier models: As of March 2026, models at or near the industry frontier include GPT-5.4, Claude Sonnet 4.6, Gemini 3.1 Pro, and Llama 4.[1]

- **Generative AI**: "Any AI system whose primary function is to generate content" (Toner 2023).
  - We typically only use the term **"generative AI" to highlight issues specific to synthetic text (which can include software code), images, video, audio, or other media**. (In other documents, "generative AI" is often used in approximately the same way that we use the term "GPAI model.")
  - Examples of generative AI: "Typical examples of generative AI systems include image generators (such as Midjourney or Stable Diffusion), large language models or multimodal models (such as GPT-4, PaLM, or Claude), code generation tools (such as [GitHub] Copilot), or audio generation tools (such as VALL-E or resemble.ai)" (Toner 2023).

- **Agentic AI:** "*Agentic AI refers to AI systems composed of* [one or more] *agents that can behave and interact autonomously in order to achieve their objectives.... Traditional software typically follows fixed pathways to solve problems. In contrast, agent-based systems* [can] *operate like independent assistants that choose and combine several actions to achieve their goals*" (GOV.UK n.d).
  - While this definition may presume a high level of autonomy, we acknowledge that AI agency exists on a spectrum of autonomy and authority (Mitchell et al. 2025, Kasirzadeh and Gabriel 2025, WEF 2024a, WEF 2025a) and cannot be viewed as binary. (For more on AI agent characteristics and properties see, Map 5.1.)

- **AI Agent:**[2] Refers to an AI system with the ability to "...make plans to achieve goals, adaptively perform tasks involving multiple steps and uncertain outcomes along the way, and interact with its environment — for example by creating files, taking actions on the web, or delegating tasks to other agents — with little to no human oversight" (Bengio et al. 2025, p. 38).

We intend our usage of the above terms to be broadly compatible with usage of the corresponding terms where applicable in the OECD classification framework (OECD 2022a, p. 64), EU AI Act (EP 2024), and the Hiroshima Process International Code of Conduct for Advanced AI Systems (G7 2023), though our focus in this document is primarily on the most broadly capable GPAI models.

---

1 Several AI companies have signed the EU GPAI Code of Practice, which includes commitments for adopting a state-of-the-art safety and security framework, risk identification, and risk analysis (EC 2025a).

2 It should be noted that our use of the term "AI agent" excludes artificial intelligence systems that exhibit no levels of autonomy, including large language models and conversational agents (chatbots), which operate without the autonomous and independent decision-making characteristics that define AI agents.

The terminology used in this document has been refined since version 1.0 to reflect focus on upstream AI "models" rather than downstream AI "systems" that incorporate a model.[3]

1.2 BACKGROUND AND PURPOSE OF THE PROFILE

GPAI models such as GPT-5.1, Gemini 3, Claude Sonnet 4.5, and Llama 4 can serve as multi-purpose AI models underpinning many end-use applications. These increasingly powerful GPAI models are the focus of cutting-edge research. They also have several qualitatively distinct properties compared to the more common, narrower machine learning models. For example, GPAI models have the potential to be applied to many sectors at once, they could lead to potential large-scale societal, environmental, security, and economic impacts, and they have emergent properties that can provide beneficial capabilities as well as lead to unexpected adverse risks[4] (Berti et al. 2025, Bommasani et al. 2021, Weidinger et al. 2021, Wei et al. 2022, Ganguli et al. 2022). Given the potential risks, developers and deployers should carry out more in-depth risk assessments — with longer time horizons and at more points in the AI system life cycle — and implement more extensive risk-mitigation controls for GPAI models than for AI with more limited capabilities.

This Profile is designed to complement the broadly applicable guidance in the NIST AI Risk-Management Framework, also known as the AI RMF (NIST 2023a), or a related AI risk-management standard, such as ISO/IEC 23894. This document provides a set of risk-management practices or controls and target outcomes specifically for identifying, analyzing, and mitigating risks of GPAI models. This guidance is meant to apply across sectors, and addresses important underlying risks and early-development risks of such technologies in a way that does not rely on great certainty about each specific end-use application. We have developed this Profile through a multi-stakeholder process, integrating input and feedback on drafts from a

3 The changes in terminology are also intended to reflect current trends in AI development and governance, in which a single large GPAI model typically plays a central role as a core part of either a general-purpose AI system (GPAIS) or a relatively narrow-purpose end-use application. (We expect that costs of applying relevant parts of the guidance in this document would typically be lower for downstream developers of GPAIS and end-use applications than for upstream developers of GPAI models; see Section 1.3 for more.) However, if highly powerful GPAIS begin to be created by combining a number of smaller models rather than relying primarily on a single core GPAI model, then we may return to focusing on GPAIS as a more inclusive term in future versions of this document.

4 In some cases, emergent properties of large-scale models could have been observed as partially emergent properties of smaller-scale models if different metrics had been used (Schaeffer et al. 2023). We believe this is an argument for working to identify capabilities and other key properties of large-scale models at an early or partially emergent stage in smaller-scale models, when responses to identified emergent properties may be more feasible and effective. For more on this, see guidance under Manage 1.3.

range of stakeholders, including organizations developing large-scale GPAI models, and other organizations across industry, civil society, academia, and government.

While this Profile focuses mainly on GPAI models, we also aim to address the need for upstream model developers' pre-release risk assessments and evaluations to consider reasonably foreseeable integrations and capabilities (e.g., tool access) of downstream AI systems incorporating a GPAI model. GPAI model providers are uniquely positioned in the value chain to anticipate and manage those risks.

GPAI model-related risk categories that we aim to address with the guidance in this document include:

- Risks stemming from the large scale and reach of GPAI models, resulting from their frequent place in the AI value chain as core models that many other systems build and rely upon;
- Risks of misuse and abuse of GPAI models, resulting from their lowering barriers for malicious activities such as generating disinformation; and
- Risks of unexpected impacts of GPAI models, resulting from the emergent behaviors, vulnerabilities, and capabilities that are often found (and continue to be found) in state-of-the-art, large-scale GPAI models.

1.3 INTENDED AUDIENCE AND USERS OF THE PROFILE

We intend for this document to be used primarily by **developers of large-scale, state-of-the-art GPAI models**; others who can benefit from use of this guidance include **downstream developers** of end-use applications or AI systems that build on a GPAI model, as well as **model evaluators and regulators**.

We believe that most AI systems could be readily identified as one of the following:

- **One of a few large-scale GPAI models**. These AI models (and especially the most broadly capable GPAI models) are the main focus of this Profile. While upstream large-scale GPAI model developers may incur governance costs (e.g., financial resources, labor, time) when employing the guidance from this Profile, they will also enjoy the benefits of risk management.

- **One of two types of AI systems that build on a GPAI model**: Such a system could be either a general-purpose AI system (GPAIS) that incorporates a GPAI model, or a relatively narrow-purpose end-use application that incorporates a GPAI model. Some aspects of these end-use applications and GPAIS are constructively addressed by parts of the guidance in this Profile. Costs (e.g., financial resources, labor, time) associated with applying the guidance in this Profile would likely be minimal to downstream developers who are building applications or GPAIS on GPAI models; only some guidance in this Profile would typically be relevant to them, and generally not the parts that would be most expensive to use.
- **One of many small-scale or standalone, narrow-purpose AI systems that do not fall under definitions for GPAI models**, and that are not within the scope of this Profile. We do not expect developers or deployers of these common AI systems to use this Profile for those AI systems.

As part of "developers of GPAI models," we aim to include all organizations and efforts that are developing such AI models, regardless of the organization size or type, and regardless of whether the organization only plans to make the AI model available to users inside the organization. (Many of the same risks, such as potential for misuse or abuse by whoever has access to the AI system, would be present to some degree for GPAI model development efforts in each of these cases.) Thus, we intend for the guidance in this document to be applicable as appropriate to:

- Open-source and open-weights GPAI model development efforts, as well as closed-source GPAI model development; and
- Research projects, and other GPAI models that a model developer does not plan to make available to users outside the organization, as well as GPAI models that a developer plans to put on the market.

1.4 BENEFITS OF THE PROFILE

1.4.1 Benefits of the Profile to Developers of GPAI Models

This Profile provides developers of GPAI models with valuable risk-management best practices that can be applied to their unique issues. For example, the Profile provides guidance on sharing of responsibilities between (a) upstream developers that create GPAI models and offer those in a manner that allows many different end uses, and (b) downstream developers that build upon the GPAI model for specific end-use applications, or who develop AI systems using upstream model provider-supplied information that may not be customized for their own application area.

This document facilitates conformity with or use of leading AI risk-management-related standards, adapting and building on the generic voluntary guidance in the NIST AI Risk-Management Framework and ISO/IEC 23894, with a focus on the unique issues faced by developers of GPAI models. It also aims to facilitate compliance with relevant regulations, such as the EU AI Act, especially for aspects related to GPAI models. For guidance on mapping to relevant clauses of ISO/IEC 23894 and the EU AI Act, see the Mapping Key Standards and Regulations to the AI Risk-Management Standards Profile for General-Purpose AI Models V1.2 (Madkour et al. 2026c).

Widespread norms for using best practices such as those detailed in this Profile can help ensure that developers of GPAI models can be competitive without compromising on practices for AI safety, security, accountability, and related issues.

1.4.2 Benefits of the Profile to Deployers, Evaluators, and Users

This Profile can provide deployers, evaluators, and users of GPAI models with increased awareness of the risks of such models, and of best practices to use in addressing those risks. This guidance can also equip deployers, evaluators, and users of GPAI models with information useful for evaluating the extent to which a model's developers have followed relevant best practices.

1.4.3 Benefits for Individuals, Society, and the Regulatory Community

Ultimately, this Profile aims to help key actors in the value chains of increasingly general-purpose AI models to achieve outcomes of maximizing benefits, and minimizing negative impacts, to individuals, communities, organizations, society, and the planet. That includes protection of fundamental rights, minimization of negative environmental impacts, and prevention of adverse events with systemic or catastrophic consequences at societal scale. There are vital relationships between principles of fairness and protecting human rights, addressing risks to individuals and groups, and addressing large-scale systemic or catastrophic risks. Some types of risks to individuals or groups can escalate to significant, severe, or catastrophic levels as they cascade across populations. Managing risks of GPAI models should include appropriate protection of human rights, and consideration of populations vulnerable to disproportionate harms. Preventing catastrophe can also be an important part of preventing unfair outcomes; often the effects of catastrophe fall disproportionately on disadvantaged people. It is critical to ensure that communities that may use or be impacted by AI systems are meaningfully involved throughout the AI lifecycle, and are given opportunities to provide feedback and report potential problems.

From a regulatory perspective, this document can be viewed as part of "soft law" norms and best practices that GPAI model developers and deployers have incentives to follow, and that regulators can consider when formulating relevant "hard law" regulations (see, e.g., Gutierrez et al. 2021).[5] We also aim to provide mapping to, and harmonization with, relevant standards (e.g., ISO/IEC 23894) and regulations (e.g., the EU AI Act), which can help to set norms for GPAI model risk-management practices and conformity across regulatory regimes.

1.5 LIMITATIONS AND CHALLENGES

This Profile has a number of limitations. Perhaps the most important is its focus on AI risk management considerations for GPAI models.. While GPAI models may be used directly in a broad range of settings, or downstream developers may create AI systems or applications that incorporate GPAI models, this Profile does not provide all guidance that might be needed in particular industry sectors or applications. This Profile also does not provide all guidance that might be needed by GPAI model developers on risk-management topics not directly related to model development and deployment, such as securing an organization's networking equipment or other information system components.

This Profile is limited to the risks that have been identified, including those that have been recorded, are emerging, or are theoretical in nature. We have based our guidance on available literature, industry best practices, stakeholder input and feedback, and our own expertise. Although some suggestions address identification of risks as they emerge, our guidance is not meant to extend past foreseeable risks. In such a rapidly evolving field, the gap between identified and total possible risks may be substantial; therefore, areas where guidance states "identify all risks" or "mitigate all identified risks" should be read with these limitations in mind. For risk categories where outcomes may be severe and irreversible — such as risks related to loss of control, large-scale manipulation, or catastrophic misuse — the gap between identified risks and total risk may be especially consequential, which may warrant additional caution beyond the specific practices outlined in this Profile.

5 As a related example, the U.S. National Telecommunications and Information Administration (NTIA) made AI accountability policy recommendations that include U.S. government procurement requirements for use of appropriate AI standards and risk-management practices such as audits. NTIA included foundation models in its considerations (NTIA 2024a).

Another limitation is the relatively nascent state of best practices for developers of GPAI models responding to these risks. We expect that best practices in this area will continue to evolve substantially, and we aim to incorporate emerging best practices in later versions of this guidance (i.e., in annual updates). As such, this guidance only represents current best practices, but does not inherently guarantee acceptable residual risk.

Users of this guidance may face challenging tradeoffs, such as the need to weigh risks and benefits, or choose between different sets of risks. One of the most challenging areas is the open-source development and release of models — or closely related release strategies, such as open-weights release — where model weights can be downloaded by outside parties. There is great value in open-source software and various forms of transparency and access to AI systems, including for helping to ensure the safety and security of an AI system's intended users. However, providing direct access to a model's weights also can increase some types of risks, including risks of malicious misuse. GPAI model developers who provide open-weights access to their models, or other developers whose model weights are leaked inadvertently, will in effect be unable to shut down or decommission AI systems that others build using those model weights. This is a consideration that should be weighed against the benefits of developing and sharing open-weights models, especially for the largest-scale and most broadly capable models that pose the greatest risks of enabling severe harms..

Many of the benefits of open-weights release, such as review and evaluation from a broader set of stakeholders and expanded access to use of a model, are possible to support either through transparency, engagement, or other openness mechanisms that do not require a model's parameter weights to become downloadable, or through the release of smaller-scale and less broadly capable open-weights models (including open-source models that can provide greater transparency than models that are only open-weights). Thus, our profile guidance includes many transparency and access provisions, including (under Govern 4.2) on reporting on the model to internal and external stakeholders (e.g., to downstream developers, regulators, users, impacted communities, etc.), e.g., using model cards, or system cards, and other transparency mechanisms. Another important part of our guidance (under Manage 2.4) is that GPAI model developers who plan to provide downloadable, fully open, or open-source access to their models should first use a staged-release approach (e.g., not releasing parameter weights until after an initial closed-source or structured-access release, when no substantial risks or harms have emerged over a sufficient time period), and should not proceed to a final step of releasing model parameter weights until a sufficient level of confidence in risk management has been established, including assessment of safety risks and risks of misuse and abuse. (That level of confidence in safety would be particularly difficult to establish for the largest-scale or most capable models, and they should be given the greatest duration and depth of pre-release evaluations, as they are the most likely to have dangerous capabilities or other emergent properties that can take some time to discover.) Our recommended approach is also consistent with recommendations on deployment of frontier models, as provided by the Partnership on AI (PAI 2023c), the Safety and Security chapter of the EU GPAI Code of Practice (EC 2025a), and others. We believe this overall approach provides actionable guidance to address some of the greatest risks to the public associated with open-sourcing powerful AI models while also providing valuable transparency[6] mechanisms, and allowing responsible open-sourcing of AI models.[7]

Effective AI risk-management also requires cohesive efforts from governments to manage certain systemic, high-impact risks that cannot be sufficiently mitigated by model vendors acting alone. Accordingly, this Profile is intended to complement, not replace, public-sector responsibilities — e.g., for setting enforceable requirements, conducting oversight, and establishing societal risk tolerances — that exceed the authority or visibility of any single developer.

6 For more detailed recommendations on GPAI risk-management transparency, documentation, and reporting, see Madkour et al. (2026b).

7 For more on risk-management tradeoffs for release strategies for frontier and near-frontier open-weights and open-source models, see, e.g., Solaiman (2023), Seger et al. (2023), PAI (2023c), Kapoor et al. (2024), Bateman et al. (2024), and NTIA (2024b).

2. Overview of Components and How to Use the Profile

2.1 BASICS

We intend for this Profile to be used in conjunction with the NIST AI RMF (NIST 2023a) and AI RMF Playbook (NIST 2023b), or an approximately equivalent set of AI risk-management guidance documents, or an AI risk-management framework or standard such as ISO/IEC 23894. In addition, we generally assume the use of appropriate guidance for risk topics not specific to AI, such as the NIST Cybersecurity Framework or ISO/IEC 27001, for broadly applicable information system security management guidance.[8]

It also can be appropriate to combine this Profile with another resource that provides supplemental guidance on particular industry sectors or applications for use-case-specific risks, metrics, and controls. This would be most appropriate for downstream developers focused on building on or applying GPAI models for specific industry sectors or use cases.

The AI RMF "core functions," or broad categories of activities, apply as appropriate across AI system lifecycles, and we provide corresponding guidance in related sections of this Profile:

- "Govern" (Section 3.1) for AI risk-management process policies, roles, and responsibilities;
- "Map" (Section 3.2) for identifying AI risks in context;
- "Measure" (Section 3.3) for rating AI trustworthiness characteristics; and
- "Manage" (Section 3.4) for decisions on prioritizing, avoiding, mitigating, or accepting AI risks.

NIST (2023a) organizes high-level functions into categories and subcategories of activities and outcomes. In addition, NIST provides more detailed guidance in a companion Playbook resource document (NIST 2023b).

Our usage of the terms "should" and "can" in the guidance in Section 3 of this document is as follows: "should" indicates our recommendation and "can" indicates something is possible.[9]

8 If specific guidance suggested in this document has become obsolete, then use analogous or related up-to-date guidance instead of, or in addition to, the older guidance.

9 This is broadly consistent with usage by ISO and other standards organizations. See, e.g., ISO (n.d.).

2.2 IMPACT AREAS, HARM FACTORS, AND TRUSTWORTHINESS CHARACTERISTICS

GPAI model-related risk topics and corresponding guidance sections in this Profile include the following topics. (Some of these topics overlap with others, in part because the guidance often involves iterative assessments for additional depth on issues identified at earlier stages.)[10]

- Reasonably foreseeable impacts (Section 3.2, Map 1.1), including impacts:
  - To individuals, including to health, safety, well-being, or fundamental rights;
  - To groups, including populations vulnerable to disproportionate adverse impacts or harms; and
  - To society, including environmental impacts.
- Significant, severe, or catastrophic harm factors (Section 3.2, Map 5.1), including:
  - Correlated bias and discrimination;
  - Impacts to societal trust or democratic processes;
  - Correlated robustness failures;
  - Capability to provide information for weaponization (e.g., CBRN, cyber);
  - Capability to manipulate or deceive humans in harmful ways (e.g., deepfakes); and
  - Loss of understanding and control of an AI system in a real-world context (e.g., the ability to escape a sandbox and replicate on another computational system).
- AI trustworthiness characteristics (Section 3.4, Measure 2), including:
  - Safety, reliability, and robustness (Measure 2.5, Measure 2.6);
  - Security and resiliency (Measure 2.7);
  - Accountability and transparency (Measure 2.8);
  - Explainability and interpretability (Measure 2.9);
  - Privacy (Measure 2.10); and
  - Fairness and bias (Measure 2.11).

Additional topics to address in future versions of the Profile are listed in Appendix 3.

10 This risk-topic structure can be seen as an example of a risk taxonomy for GPAI models, though the topics listed here are not collectively exhaustive and mutually exclusive. For other risk taxonomies, see, e.g., Footnote 5 in the NIST Generative AI Profile, NIST AI 600-1 (Autio et al. 2024), Section 2 of the International Scientific Report on the Safety of Advanced AI (Bengio et al. 2025), TASRA: a Taxonomy and Analysis of Societal-Scale Risks from AI (Critch and Russell 2023), and the MIT AI Risk Repository (MIT n.d.a).

2.2.1 Risk Taxonomies

Several resources have emerged in recent years that provide useful categorization of AI risks (MIT n.d.a, Bengio et al. 2025, Autio et al. 2024, and EC 2025a).

Each of these resources utilizes a different approach to categorizing AI risks. The MIT Risk Repository adopts a domain-based structure to organize risks into seven categories: discrimination and toxicity; privacy and security; misinformation; malicious actors; human-computer interaction; socioeconomic and environmental; and AI system safety, failures, and limitations. In contrast, the International AI Safety Report aggregates risks into three broad classes that foreground the pathways through which they may materialize, grouping risks by whether they arise from malicious use, malfunctions, or systemic issues. The NIST Generative AI Profile outlines 12 risks unique to generative AI (e.g., confabulation, information security, data privacy, abusive content). The EU GPAI Code of Practices introduces a multi-dimensional framework that maps risks by affected social interests (e.g., risks to public health, safety, and fundamental rights), the source of risk (e.g., model capabilities, model propensities, and model affordances), and a specified list of systemic risks, including CBRN, loss of control, cyber offense, and harmful manipulation.

For clarity, we have compiled a consolidated, Profile-specific organization of risks that draws from all four sources. This compilation is intended solely to serve the analytical needs and scope of the GPAI Profile and may be refined in subsequent versions.

Bias, Discrimination, and Toxicity
- Unfair discrimination and misrepresentation
- Toxicity (dangerous, violent, or hateful content)

Privacy and Security
- Security vulnerabilities
- Data privacy violations

Malicious Use and Misuse
- CBRN weapons
- Cyber offense
- Advanced persuasion and manipulation
- Fraud and scams
- Disinformation
- Deep fakes
- Generation of obscene, degrading, and/or abusive content

  - AI-generated child sexual abuse material (CSAM) and non-concentual intimate imigary (NCII)
  - AI-facilitated nonconsensual sexual abuse and exploitation
- Surveillance

Human-Computer Interaction

- Overreliance
- Anthropomorphic AI and psychological dependence
- Loss of control

Socioeconomic and Ecosystem Harms

- Misinformation
- Pollution of the information ecosystem
- Loss of consensus reality
- Power disparities in inequality
- Collective disempowerment
- Economic and labor market disruptions
- Environmental harms
- Global AI R&D divide
- Market concentration
- Infringement of fundamental rights
- Infringement of intellectual property rights

Failures, Limitations, and Misalignment

- Self-proliferation
- Self-modification
- Self-exfiltration
- Self-replication
- Evaluation deception
- Reward hacking
- Hallucinations, confabulation, and lack of performance reliability
- Goal misalignment
- Evasion of human oversight

The GPAI Profile guidance is largely structured around impact areas, harm factors, and trustworthiness characteristics (listed under section 2.2). The risk taxonomy summarized here complements this approach by helping readers map specific risks and the broader harms they may produce, particularly where multiple risks contribute to second-order effects. Certain impacts, like the erosion of societal trust or democratic processes — or risks to public health and security, safety, fundamental rights, or to society as a whole — may be considered second-order effects, or consequences of a combination of other risks. For instance, impacts to societal trust or democratic processes may arise from a combination of disinformation, overreliance, pollution of the information ecosystem, and deepfakes.

Future iterations of the Profile[11] may refine these subcategories, restructure them based on a more developed taxonomy, or adopt an existing authoritative framework if a stronger consensus emerges. This iterative approach ensures the Profile can evolve with stakeholder feedback, emerging research, and the latest developments in the AI risk landscape.

2.3 HIGH-PRIORITY RISK-MANAGEMENT STEPS AND PROFILE GUIDANCE SECTIONS

Users of this Profile should place high priority on the following risk-management steps and corresponding guidance sections.[12] (Appropriately applying the Profile guidance for the following steps should be regarded as the baseline or minimum expectations; users of this Profile can exceed the minimum expectations by also applying guidance in other sections.)

- **Check or update, and incorporate, each of the following high-priority risk-management steps when making go/no-go decisions,** especially on whether to proceed on major stages or investments for development or deployment of cutting-edge large-scale GPAI models (Manage 1.1).
- **Take responsibility for risk assessment and risk-management tasks for which your organization has access to information, possesses requisite resources, or has the opportunity to develop capabilities sufficient for constructive action, especially when these are substantially greater than that of others in the value chain** (Govern 2.1).
  - We also recommend applying this principle throughout other risk assessment and risk-management steps, and we refer to it frequently in other guidance sections.
- **Set risk-tolerance thresholds to prevent unacceptable risks** (Map 1.5).
  - For example, the NIST AI RMF 1.0 recommends the following: "In cases where an AI system presents unacceptable negative risk levels — such as where significant negative

11 For more on plans for future iterations of the GPAI Profile see, Appendix 3.

12 It also can be appropriate to follow the guidance in this document for these risk-management steps, but to apply and document them under other, closely related risk-management steps (typically noted in this document with "see also" statements pointing to guidance in other sections of the Profile). For example, if your organization sets risk-tolerance thresholds under Govern 1.3 instead of under Map 1.5, then as part of your organization's process for Govern 1.3, it can be appropriate to follow the guidance under Map 1.5.

impacts are imminent, severe harms are actually occurring, or catastrophic risks are present — development and deployment should cease in a safe manner until risks can be sufficiently managed" (NIST 2023a, p. 8).[13]

- **Collect, consider, and integrate feedback from external stakeholders and third-party evaluators** (Govern 5.1).
    - Prioritize partnerships with independent third-party evaluators and auditors throughout the model lifecycle, particularly when assessing the model for dangerous capabilities, and establish clear and accessible feedback channels for users or impacted persons.
- **Identify reasonably foreseeable uses, misuses, and abuses for a GPAI model** (e.g., automated generation of toxic or illegal content or disinformation, or aiding with proliferation of cyber, chemical, biological, or radiological weapons), and identify reasonably foreseeable potential impacts (e.g., to fundamental rights) (Map 1.1).
- **Identify whether a GPAI model could lead to significant, severe, or catastrophic impacts,** e.g., because of correlated failures or errors across high-stakes deployment domains, dangerous emergent behaviors or vulnerabilities, or harmful misuses and abuses (Map 5.1).
- **Use red teams and adversarial testing** as part of extensive interaction with GPAI models to identify dangerous capabilities, vulnerabilities, or other emergent properties of such systems (Measure 1.1).
- **Track important identified risks** (e.g., vulnerabilities from data poisoning and other attacks, or misspecification of objectives), even if they cannot yet be measured (Measure 1.1 and Measure 3.2).
- **Implement risk-reduction controls as appropriate** throughout a GPAI model's lifecycle, e.g., through independent auditing, incremental scale-up, red-teaming, and other steps (Manage 1.3, Manage 2.3, and Manage 2.4).
- **Incorporate identified AI system risk factors — and circumstances that could result in impacts or harms — into reporting and engagement with internal and external stakeholders** (e.g., with downstream developers, regulators, users, impacted communities, etc.), e.g., using model cards, system cards, and other transparency mechanisms (Govern 4.2).
- **Continuously monitor the model post-deployment** by actively gathering relevant information for risk evaluation, and implement an incident response plan (Manage 4.1).

13 Risk tolerances should additionally account for uncertainty and gaps in current risk identification methods, not only concretely identified risk sources.

We also recommend **documenting the process used in considering risk mitigation controls, the options considered, and reasons for the choices made**. Documentation on many items should be shared in publicly available materials such as system cards. Some details on particular items, such as security vulnerabilities, can be responsibly omitted from public materials to reduce misuse potential, especially if available to auditors, Information Sharing and Analysis Organizations, or other parties as appropriate.

Refer to our **General-Purpose AI Risk-Management Transparency, Documentation, and Reporting Recommendations** document (Madkour et al. 2026b) for recommendations on transparency, documentation, and reporting that are in alignment with the GPAI Profile, as well as for crosswalks between leading AI transparency standards and governance resources, including the National Institute of Standards and Technology (NIST) AI Risk-Management Framework (AI RMF) Playbook (NIST 2023b), the G7 Hiroshima AI Process (HAIP) reporting framework (OECD.AI 2025), and the EU General-Purpose AI (GPAI) Code of Practice (CoP) (EC 2025a).

3. Guidance

The tables in this section provide applicability of NIST AI RMF categories and subcategories, and supplemental guidance, for GPAI models. The tables address the following AI RMF functions: Table 1 for Govern, Table 2 for Map, Table 3 for Measure, and Table 4 for Manage.

Broadly speaking, all areas of current NIST AI RMF guidance (NIST 2023a, 2023b) are at least partly applicable for GPAI models. However, the activities and outcomes for some categories or subcategories seem to be of higher priority than others for general purpose models. In this section, we include a number of excerpts from the NIST AI RMF Playbook (NIST 2023b) that seem particularly valuable for GPAI model developers, given typical model architectures and development practices currently in use. These excerpts are in *italic font*, and are preceded by statements of the form, "In the NIST AI RMF Playbook guidance for __, particularly valuable action and documentation items for GPAI models include __." When particularly applicable, we have also included excerpts from the NIST Generative AI Profile, NIST AI 600-1 (Autio et al. 2024). These are also *in italic font* and preceded by statements of the form, "In the NIST GAI Profile guidance for __, additional particularly valuable actions include __."[14] Sub-categories that are in the NIST Generative AI Profile, but do not include excerpts, have the NIST Generative AI Profile listed as a resource.

3.1 GUIDANCE FOR NIST AI RMF GOVERN SUBCATEGORIES

Table 1: Guidance for NIST AI RMF Govern Subcategories

Applicability and Supplemental Guidance for GPAI Models	Resources
Govern 1: Policies, processes, procedures, and practices across the organization related to the mapping, measuring, and managing of AI risks are in place, transparent, and implemented effectively.	
Govern 1.1: Legal and regulatory requirements involving AI are understood, managed, and documented.	
The legal and regulatory environment for GPAI models is evolving quickly and will require regular assessment for continued compliance. GPAI model developers, deployers, and users should assess the extent to which their activities would fall under GPAI model-related laws or regulations. (See, e.g., policy trackers such as OECD.AI n.d.a and IAPP 2024, 2025.)	NIST AI RMF Playbook (NIST 2023b) NIST Generative AI Profile, NIST AI 600-1 (Autio et al. 2024)

14 While most of these NIST Generative AI Profile excerpts are included verbatim, some have been summarized to reduce their length.

Applicability and Supplemental Guidance for GPAI Models	Resources
• The EU AI Act entered into force on August 1, 2024 and applies to any AI product or service offered in the EU regardless of where it was established. The EU AI Act includes Obligations for Providers of General-Purpose AI Models in Article 53, as well as Obligations for Providers of General-Purpose AI Models with Systemic Risk (i.e., frontier or near-frontier models) in Article 55. ◦ Additionally, signatories of the EU GPAI Code of Practice (CoP) must comply with the commitments and measures set forth in the Safety and Security, Transparency, and Copyright chapters of the GPAI CoP (EC 2025a). • California's Transparency in Frontier Artificial Intelligence Act, SB53 (California Legislature 2025) took effect in January 2026 and sets requirements for model developers of advanced AI systems to publish safety frameworks, report critical incidents, and provide whistleblower protections. • The Texas Responsible Artificial Intelligence Governance Act (TRAIGA) HB149 has been signed into law and includes requirements around prohibited harmful uses of AI and disclosure to consumers (Texas Legislature 2025). Many pre-existing legal and regulatory requirements — for example those related to copyright, data protection, discrimination, and privacy rights — are particularly relevant to GPAI models that are trained on large swaths of the internet. International human rights law is also highly relevant to GPAI models; see, e.g., The Risk-Management Profile for Artificial Intelligence and Human Rights (DOS 2024). See the Mapping Key Standards and Regulations to the AI Risk-Management Standards Profile for General-Purpose AI Models V1.2 (Madkour et al. 2026c) supporting document for mappings of our guidance to key standards, commitments, and regulations, including ISO/IEC 23894, the White House AI Commitments, and the EU AI Act.	For policy tracking: OECD National AI Policies & Strategies Repository (OECD. AI n.d.a) Global AI Law and Policy Tracker (IAPP 2024) US State Governance Legislation Tracker (IAPP 2025) On the EU AI Act: EP (2024) On copyright and fair use: Henderson et al. (2023) Samuelson (2023) For data audits and copyright filtering: C4 (Dodge et al. 2021), see also (Birhane et al. 2021) EU GPAI Code of Practice (EC 2025a) On human rights obligations: Risk-Management Profile for Artificial Intelligence and Human Rights (DOS 2024)
Govern 1.2: The characteristics of trustworthy AI are integrated into organizational policies, processes, procedures, and practices.	
The characteristics of trustworthy AI, described in the NIST AI RMF, include: valid and reliable, safe, secure and resilient, accountable and transparent, explainable and interpretable, privacy-enhanced, and fair with harmful bias managed. For GPAI models, there are some unique or particularly important considerations related to ensuring the characteristics of trustworthy AI are integrated into organizational policies, processes, procedures, and practices (Newman 2023, Wang et al. 2023). Some of these are mentioned below; see also the more detailed considerations and guidance throughout this document:	NIST AI RMF Playbook (NIST 2023b) NIST Generative AI Profile, NIST AI 600-1 (Autio et al. 2024) Newman (2023) Wang, Chen et al. (2023) Kaur et al (2022) Brundage et al. (2020)

<table>
<tr><th>Applicability and Supplemental Guidance for GPAI Models</th><th>Resources</th></tr>
<tr><td>Valid and Reliable:
• E.g., improve predictability, enable verification, review dependencies on external parties, assess quality of training data, and train operators of the system to exercise oversight and avoid overconfidence in the system.
Safe:
• E.g., establish risk thresholds to enable proactive governance, design reliable technical and procedural controls, re-evaluate safety regularly, assess shifts over time, report incidents and adverse impacts, and institute processes for system recall, retirement, and/or decommissioning.
Fair with Harmful Bias Managed:
• E.g., engage with impacted communities, test for biased or discriminatory outputs, review impacts on human rights and wellbeing, assess accessibility of user interface, use balanced datasets to train models, and determine how to equitably distribute benefits.
Secure and Resilient:
• E.g., assess robustness in novel environments, establish protections against adversarial attacks, and establish a coordinated policy to encourage responsible vulnerability research and disclosure.
Explainable and Interpretable:
• E.g., ensure users know how to interpret system behavior and outputs, including limitations.
Privacy-Enhanced:
• E.g., enable people to consent to the uses of their data and opt out of the uses of their data, and notify users about privacy and security breaches.
Accountable and Transparent:
• E.g., establish transparency and documentation policies and processes, determine a publication/release strategy, inform users when they are interacting with the AI system or viewing AI-generated content, allow people to opt out, support independent third-party auditing and evaluation, and provide redress to people who are negatively affected.
Responsible:
• E.g., review, assess, and potentially control for adverse impacts on global financial systems, supply chains, markets, labor markets, the environment, and natural resources.

For Govern 1.2 guidance specific to agentic AI and AI agents, see our Risk-Management Standards Profile for Agentic AI (Madkour et al. 2026a).</td><td>On secure software and AI development approaches:
CISA (2023)
NIST SP 800-218 (Souppaya et al. 2022)
NIST SP 800-218A (Booth et al. 2024)
Bai et al. (2022)</td></tr>
<tr><td colspan="2">Govern 1.3: Processes, procedures, and practices are in place to determine the needed level of risk-management activities based on the organization’s risk tolerance.</td></tr>
<tr><td>GPAI models can have greater impacts or pose greater risks than smaller or less capable AI systems due to their potential use in many different downstream applications. Therefore, for GPAI models, it is often appropriate to make model risk assessment and management a higher priority, and devote more resources, as compared with lower-capability and lower-impact AI systems.

When defining risk tiers, organizations should establish clear measurable categories based on model capabilities, as well as metrics such as model propensities, risk estimates, and others (see, Measure 4.1 for further criteria and Measure 7.3 for documentation practices in the Safety and Security chapter of the EU GPAI Code of Practice (EC 2025a)).</td><td>NIST AI RMF Playbook (NIST 2023b)
NIST Generative AI Profile, NIST AI 600-1 (Autio et al. 2024)
EU GPAI Code of Practice (EC 2025a)</td></tr>
</table>

<table>
<tr><th>Applicability and Supplemental Guidance for GPAI Models</th><th>Resources</th></tr>
<tr><td>For guidance on determining and documenting organizational risk tolerance, see material in this document under Map 1.5.

In the NIST GAI Profile guidance for Govern 1.3, additional particularly valuable actions include:
• Define risk tiers.
• Establish minimum thresholds as part of deployment approval.
• Establish a plan to periodically evaluate whether the model may misuse CBRN information or capabilities and/or offensive cyber capabilities.
• Obtain input from stakeholder communities on unacceptables uses.
• Maintain an updated hierarchy of identified and expected GAI risks connected to contexts of GAI model advancement and use.
• Reevaluate organizational risk tolerances to account for unacceptable negative risk.
• Devise a plan to halt development or deployment of a GAI system that poses unacceptable negative risk.

See also the material in this document under Map 1.1 and Map 5.1 for related guidance on GPAI model impact assessment, including on impact identification and impact magnitude rating, and under Map 1.5 on risk tolerance, including on setting unacceptable-risk thresholds.</td><td>For risk tier considerations see:
• GV-1.3-001 in NIST AI 600-1 (Autio et al. 2024)
• Caputo et al. (2025)</td></tr>
<tr><td colspan="2">Govern 1.4: The risk-management process and its outcomes are established through transparent policies, procedures, and other controls based on organizational risk priorities.</td></tr>
<tr><td>Establish, and make available and easily accessible, relevant public-facing policies, including:
• Privacy policies (e.g., Anthropic 2025a);
• Acceptable use policies (AUPs), or terms of use policies (e.g., OpenAI 2024c);
• Prohibited use policies (e.g., Google 2024a); and
• Data use and retention policies.

Traditional incident response plans are often limited to cybersecurity incidents (e.g., data breaches, unauthorized access). However, AI incidents extend beyond conventional cybersecurity boundaries and include risks specific to machine learning (e.g., model extraction, data poisoning, or adversarial attacks). Organizations must establish and maintain AI incident response plans specifically tailored to address the threat landscape of AI systems (Leong and Atherton 2023).
The incident response plan should clearly define:
• Incident classification criteria, including incident severity levels (if applicable);
• Any other organizational policies that apply in the event of an incident;
• Roles and responsibilities, including who is responsible for making go/no-go decisions to pause or stop operations; and
• Reporting and notification procedures (e.g., notification of oversight bodies and relevant stakeholders).
The incident response plan should contain six stages:
• Preparation: design the policies and procedures that outline operations, roles and responsibilities, definitions, thresholds, foreseeable risks, and training.
• Identification: procedures for detecting and verifying risk, monitoring models, feedback channels, documentation, and information sharing.</td><td>NIST AI RMF Playbook (NIST 2023b)
NIST Generative AI Profile, NIST AI 600-1 (Autio et al. 2024)
Bender et al. (2022)
Gebru et al. (2021)
Mitchell et al. (2019)

Common Elements of Frontier AI Safety Policies (METR 2025)
AI incident response plans (Leong and Atherton 2023)

List of frontier AI safety policies (METR n.d.a)</td></tr>
</table>

Applicability and Supplemental Guidance for GPAI Models	Resources

• **Containment:** respond to damage, stop or pause operations if needed, implement backup as necessary, begin documentation, and apply mitigations.
• **Eradication:** remove systems until mitigations are implemented and verified through further testing, including assessing any effects on upstream or downstream dependencies.
• **Recovery:** revised or replacement model is hardened before deployment, and performance metrics are updated.
• **Lessons Learned:** review incident documentation and create summarized reports for the incident outcomes and responses, including documenting any gaps. Share documentation with appropriate teams and stakeholders, and update training and testing procedures as appropriate.
For more on AI incident response plans, see Leong and Atherton (2023).
 	
In the NIST AI RMF Playbook guidance for Govern 1.4, particularly valuable action and documentation items for GPAI models include:
• *Establish and regularly review documentation policies that, among others, address information related to:*
  ◦ *Expected and potential risks and impacts*
  ◦ *Assumptions and limitations*
  ◦ *Description and characterization of training data*
  ◦ *Testing and validation results (including explanatory visualizations and information)*
  ◦ *Down- and up-stream dependencies*
  ◦ *Plans for deployment, monitoring, and change management*
  ◦ *Stakeholder engagement plans*
• *Establish policies and processes regarding public disclosure of the use of AI and risk-management material such as impact assessments, audits, model documentation and validation, and testing results.*
• *Document and review the use and efficacy of different types of transparency tools and follow industry standards at the time a model is in use.*
 	
In the NIST GAI Profile guidance for Govern 1.4, additional particularly valuable actions include:
• *Establish policies and mechanisms to prevent GAI systems from generating CSAM, NCII, or content that violates the law.*
• *Establish transparent acceptable use policies for GAI that address illegal use or applications of GAI.*
 	
(When considering disclosure of risk-management material, such as impact assessments, audits, model documentation and validation and testing results, see also the material under Govern 4.2 for related guidance on documentation and communication.)
 	
For Govern 1.4 guidance specific to agentic AI and AI agents, see our Risk-Management Standards Profile for Agentic AI (Madkour et al. 2026a).	

<table>
<tr><th>Applicability and Supplemental Guidance for GPAI Models</th><th>Resources</th></tr>
<tr><td colspan="2">Govern 1.5: Ongoing monitoring and periodic review of the risk-management process and its outcomes are planned and organizational roles and responsibilities clearly defined, including determining the frequency of periodic review.</td></tr>
<tr><td>
Plan to identify GPAI model impacts (including to human rights) and risks (including potential uses, misuses, and abuses), starting from an early AI lifecycle stage and repeatedly through new lifecycle phases, or as new information becomes available. This is particularly important for GPAI models, which can have large numbers of uses, risks, and impacts, including from emergent capabilities and vulnerabilities.

On GPAI model lifecycle and when to assess risks:

For larger machine learning models, iterations are often slower than typical agile sprints. The typical pipeline for larger models is to pre-train a model, analyze it, customize it, reanalyze, apply further customizations, then deploy, monitor, and eventually decommission.. (Here we use “analyze” as a shorthand for probing, stress-testing, red-teaming, monitoring in simulated environments, etc.)
On red-teaming, see e.g., Ganguli, Lovitt et al. (2022), and guidance in this document under Measure 1.1.

All the relevant parties involved in the R&D process, especially researchers, should have some minimal knowledge on the risks of GPAI models or be taught about such risks upon their inclusion on an advisory team.
For larger models or close-to-frontier models, “Map” activities to identify risks should also happen after model training to incorporate developer findings about a model’s capabilities.
On identifying potential uses, misuses, and abuses of a GPAI model:

Identify potential use cases during early stages of the AI system lifecycle, such as the plan and design stages, at minimum.
Identify misuse or abuse cases during all major stages of the AI system lifecycle (or approximate equivalents in Agile/iterative development sprints), such as: plan, data collection, design, train/build/buy, test and evaluation, deploy, operate and monitor, and decommission.
Revisit use and misuse case identification at key intended milestones, or at periodic intervals (e.g., at least annually), whichever comes first.
Create a plan for ongoing use case identification and categorization to extend identified uses, misuses, and abuses, based on information gained continuously from sources such as:

Downstream user and developer exploration of the AI system; and
API misuse and abuse monitoring.

When making go/no-go decisions, especially on whether to proceed on major stages or investments for development or deployment of frontier models, see guidance in this document under Manage 1.1.

It can be valuable to revisit risk assessment at these intervals, especially prior to beginning a new frontier model training run. At or near a foundation model frontier, it is particularly important to obtain and integrate new information on emergent properties of frontier models before committing to the next high-cost, large-scale model training effort.

Review of the risk-management process should be conducted periodically (e.g., every six months), or if:

There are reasonable grounds to believe that the adequacy of the risk-management process has been, or may soon be, materially compromised;

</td><td>Barrett et al. (2022)
PAI (2023a)
NIST AI RMF Playbook (NIST 2023b)
NIST Generative AI Profile, NIST AI 600-1 (Autio et al. 2024)
EU GPAI Code of Practice (EC 2025a)
Grotto and Dempsey (2021)</td></tr>
</table>

Applicability and Supplemental Guidance for GPAI Models	Resources
◦ Serious incidents or near misses involving the model, or similar models, occur that suggest risks may be unacceptable; ◦ Risks from the model, or similar models, have changed or are likely to change significantly; and ◦ Assessment of risk-management practices or frameworks should include an assessment of adequacy (i.e., adequacy in addressing risks stemming from the model), and adherence to internal governance processes. • For additional information, see Measure 1.3 in the Safety and Security chapter of the EU GPAI Code of Practice (EC 2025a). In the NIST AI RMF Playbook guidance for Govern 1.5, particularly valuable action and documentation items for GPAI models include: • *Establish policies to allocate appropriate resources and capacity for assessing impacts of AI systems on individuals, communities, and society.* • *Establish policies and procedures for monitoring and addressing AI system performance and trustworthiness, including bias and security problems, across the lifecycle of the system.* • *Establish policies for AI system incident response, or confirm that existing incident response policies apply to AI systems.* • *Establish policies to define organizational functions and personnel responsible for AI system monitoring and incident response activities.* • *Establish mechanisms to enable the sharing of feedback from impacted individuals or communities about negative impacts from AI systems.* • *Establish mechanisms to provide recourse for impacted individuals or communities to contest problematic AI system outcomes.* In the NIST GAI Profile guidance for Govern 1.5, additional particularly valuable actions include: • *Define organizational responsibilities for periodic review of content provenance and incident monitoring for GAI systems.* • *Establish organizational policies and procedures for after-action reviews of GAI system incident response and incident disclosures, to identify gaps; update incident response and incident disclosure processes as required.* • *Maintain a document retention policy to keep history for test, evaluation, validation, and verification (TEVV), and digital content transparency methods for GAI.* For Govern 1.5 guidance specific to agentic AI and AI agents, see our Risk-Management Standards Profile for Agentic AI (Madkour et al. 2026a).	
Govern 1.6: Mechanisms are in place to inventory AI systems and are resourced according to organizational risk priorities.	
In the NIST GAI Profile guidance for Govern 1.6, additional particularly valuable actions include: • *Enumerate organizational GAI systems for incorporation into AI system inventory and adjust AI system inventory requirements to account for GAI risks.* • *In addition to general model, governance, and risk information, consider the following items in GAI system inventory entries: Data provenance information (e.g., source, signatures, versioning, watermarks); Known issues reported from internal bug tracking or external information sharing resources (e.g., AI incident database, AVID, CVE, NVD, or OECD AI incident monitor); Human oversight roles and responsibilities; Special rights and considerations for intellectual property, licensed works, or personal, privileged, proprietary or sensitive data; Underlying foundation models, versions of underlying models, and access modes.*	NIST AI RMF Playbook (NIST 2023b) NIST Generative AI Profile, NIST AI 600-1 (Autio et al. 2024)

<table>
<tr><th>Applicability and Supplemental Guidance for GPAI Models</th><th>Resources</th></tr>
<tr><td colspan="2">Govern 1.7: Processes and procedures are in place for decommissioning and phasing out AI systems safely and in a manner that does not increase risks or decrease the organization’s trustworthiness.</td></tr>
<tr><td>GPAI model developers that publicly release the model parameter weights for their models, and model developers that suffer a leak of model weights, will in effect be unable to decommission AI systems that others build using those model weights. (See also guidance in this document under Manage 2.4, recommending structured access or staged release approaches, including for GPAI model developers that plan on open-weights or open-source releases of their models.)

In the NIST AI RMF Playbook guidance for Govern 1.7, particularly valuable action and documentation items for GPAI models include:
• Establish policies for decommissioning AI systems. Such policies typically address:
◦ User and community concerns, and reputational risks.
◦ Business continuity and financial risks.
◦ Up and downstream system dependencies.
◦ Regulatory requirements (e.g., data retention).
◦ Potential future legal, regulatory, security, or forensic investigations.
◦ Migration to the replacement system, if appropriate.
• If anyone believes that the AI no longer meets this ethical framework, who will be responsible for receiving the concern and as appropriate investigating and remediating the issue? Do they have authority to modify, limit, or stop the use of the AI?

In the NIST GAI Profile guidance for Govern 1.7, additional particularly valuable actions include:
• Consider the following factors when decommissioning GAI systems: Data retention requirements; Data security, e.g., containment, protocols, Data leakage after decommissioning; Dependencies between upstream, downstream, or other data, internet of things (IOT) or AI systems; Use of open-source data or models; Users’ emotional entanglement with GAI functions.

For Govern 1.7 guidance specific to agentic AI and AI agents, see our Risk-Management Standards Profile for Agentic AI (Madkour et al. 2026a).</td><td>NIST AI RMF Playbook (NIST 2023b)
NIST Generative AI Profile, NIST AI 600-1 (Autio et al. 2024)</td></tr>
<tr><td colspan="2">Govern 2: Accountability structures are in place so that the appropriate teams and individuals are empowered, responsible, and trained for mapping, measuring, and managing AI risks.</td></tr>
<tr><td colspan="2">Govern 2.1: Roles and responsibilities and lines of communication related to mapping, measuring, and managing AI risks are documented and are clear to individuals and teams throughout the organization.</td></tr>
<tr><td>Clearly allocate responsibilities for systemic risk oversight, ownership, and assurance across governance levels to ensure effective and accountable responses (EC 2025a, Measure 8.1).
• This includes clearly defining roles and responsibilities for monitoring, periodic reviews, and updates to the risk-management process and its outcomes.

Regarding roles and responsibilities across a GPAI model value chain:
• Take responsibility for risk assessment and risk-management tasks for which your organization has access to information, possesses requisite resources, or has the opportunity to develop capabilities sufficient for constructive action, especially when these are substantially greater than of others in the value chain, such as:</td><td>Barrett et al. (2022)
NIST AI RMF Playbook (NIST 2023b)
NIST Generative AI Profile, NIST AI 600-1 (Autio et al. 2024)
EU GPAI Code of Practice (EC 2025a)
PAI (2023c)
Schuett (2022)
Srikumar et al. (2024)</td></tr>
</table>

Applicability and Supplemental Guidance for GPAI Models	Resources
◦ Identifying, categorizing, and describing risks by using risk taxonomies (e.g., Slattery et al. 2025, Weidinger et al. 2022), engaging with experts, identifying threat scenarios, and carrying out scenario analysis (Li, Ren et al. 2022). ◦ Assessing and mitigating early-stage development risks, including for AI research projects and AI systems that the organization does not plan to make available to others. ◦ Testing and documentation that require direct access to training data or the AI model, such as on knowledge limits and dangerous capabilities. ◦ Employing responsible data collection and pre-processing practices that support effective risk management, mitigation, and monitoring. ◦ Identifying reasonably foreseeable uses, misuses, and abuses of the AI system. ◦ Implementing appropriate precautions to prevent or mitigate identified potential misuses or abuses.[15] ◦ Explicitly articulating organizational risk tolerances and setting thresholds for intolerable risks to inform decisions on further model development and deployment. ◦ Making necessary information available to stakeholders, including downstream developers and deployers building on base models, and independent auditors or others as appropriate (e.g., to enable third-party auditability). » Make as much information available on AI risk factors, incidents (including near-miss incidents), knowledge limits, etc., as reasonably possible to all audiences.[16] » Provide additional information to downstream and end-use application developers and deployers as appropriate to meet their risk-management needs. » Make sure that third-party or external evaluators and/or auditors are included in pre- and post-deployment TEVV processes. • **Downstream developers and deployers** of end-use applications built on GPAI models **should be responsible for risk assessment and risk-management tasks for which which the organization has access to information, possesses requisite resources, or has the opportunity to develop capabilities sufficient for constructive action, especially when these are substantially greater than of others in the value chain**,, such as: ◦ Establishing specific context for their intended end-use application(s), and applying risk-management processes appropriate for that specific context. ◦ Utilizing information provided by the upstream provider of a GPAI model, and requesting additional information as needed. ◦ Reporting to the upstream provider, and considering reporting to others — such as information sharing and analysis organizations (ISAOs) or regulators as appropriate — any critical GPAI model vulnerabilities, biases, incidents (including near-miss incidents), etc., that would have high impacts on other downstream developers or deployers. ◦ Documenting and reporting system failures, incidents, near-misses, and system recall decisions to relevant stakeholders in a timely manner.	AI Organizational Responsibilities — Governance, risk-management, Compliance and Cultural Aspects (CSA 2024)

15 See also Manage 1.3 guidance on defining and communicating to key stakeholders whether any potential use cases would be disallowed/unacceptable, as well as Manage 2.4 guidance on staged releases and structured access for frontier and near-frontier models.

16 See also guidance in this document under Govern 4.2 and Govern 4.3 on information to share, and see Section 3.4.2.1 of Barrett et al. (2022) for guidance on providing stakeholders information on reasonably foreseeable risks without providing adversaries too much information.

Applicability and Supplemental Guidance for GPAI Models	Resources
• Downstream developers and deployers extending GPAI models (e.g., via fine-tuning training on data curated by the downstream developer) should also consider applying guidance for upstream developers (e.g., on testing and documentation that require direct access to fine-tuning training data) for any substantial extensions of the underlying base model. Fine-tuned versions of the underlying base models often have capabilities that base models do not. Drawing clear lines to demarcate the different types of responsibilities to map to different roles can be a useful first step to structuring organizational risk-management culture. Responsibilities can be differentiated into risk oversight, risk ownership, risk support and monitoring, and risk assurance. • Oversight should rest with a supervisory committee or independent body, while executive members responsible for risk-generating activities (e.g., research or product development) should hold ownership, delegating as needed to operational managers. • Monitoring functions should be assigned to executives independent from core business activities, such as a Chief Risk Officer, and assurance should be provided through internal audit functions or external reviewers. Such a structured allocation ensures accountability, independence, and proportionality to organizational complexity. For additional information, see Measure 8.1 in the Safety and Security chapter of the EU GPAI Code of Practice (EC 2025a). Alternatively, consider implementing "Three Lines of Defense," or 3LoD (Schuett 2023, Robinson and Ginns 2024), when defining roles and responsibilities for risk governance accountability within a single GPAI model developer or deployer organization: • Roles can include: 1. Research team as the first line, ultimately the head of research or equivalent. 2. Risk-management team as the second line, ultimately the chief risk officer (CRO) or equivalent; this can also include the legal and compliance team, technical safety team, and security team. 3. Internal audit as third line, ultimately the chief audit executive (CAE); this can also include the ethics board. • Illustrative reporting responsibilities can include: 1. The first line reports to the CEO. 2. The second line reports to the CEO; and the CRO reports to the board risk committee. 3. The third line reports to the board of directors or the board audit committee; the CAE is often part of the board audit committee. When dealing with agentic AI, develop effective human-agentic AI management hierarchies that preserve human authority. In the NIST AI RMF Playbook guidance for Govern 2.1, particularly valuable action and documentation items for GPAI models include: • *Establish policies that define the AI risk-management roles and responsibilities for positions directly and indirectly related to AI systems, including, but not limited to - Boards of directors or advisory committees - Senior management - AI audit functions - Product management - Project management - AI design - AI development - Human-AI interaction - AI testing and evaluation - AI acquisition and procurement - Impact assessment functions - Oversight functions.*	

<table>
<tr><th>Applicability and Supplemental Guidance for GPAI Models</th><th>Resources</th></tr>
<tr><td>Establish policies that promote regular communication among AI actors participating in AI risk-management efforts.
Establish policies that separate management of AI system development functions from AI system testing functions, to enable independent course-correction of AI systems.
In the NIST GAI Profile guidance for Govern 2.1, additional valuable actions include:
When systems may raise national security risks, involve national security professionals in mapping, measuring, and managing those risks.
Create mechanisms to provide protections for whistleblowers who report, based on reasonable belief, when the organization violates relevant laws or poses a specific and empirically well-substantiated negative risk to public safety (or has already caused harm).
For Govern 2.1 guidance specific to agentic AI and AI agents, see our Risk-Management Standards Profile for Agentic AI (Madkour et al. 2026a). For Govern 2.1 transparency recommendations, see our General-Purpose AI Risk-Management Transparency, Documentation, and Reporting Recommendations document (Madkour et al. 2026b).</td><td></td></tr>
<tr><td colspan="2">Govern 2.2: The organization’s personnel and partners receive AI risk-management training to enable them to perform their duties and responsibilities consistent with related policies, procedures, and agreements.</td></tr>
<tr><td>In the NIST AI RMF Playbook guidance for Govern 2.2, particularly valuable action and documentation items for GPAI models include:
Ensure that trainings comprehensively address technical and socio-technical aspects of AI risk-management.
Define paths along internal and external chains of accountability to escalate risk concerns.</td><td>NIST AI RMF Playbook (NIST 2023b)</td></tr>
<tr><td colspan="2">Govern 2.3: Executive leadership of the organization takes responsibility for decisions about risks associated with AI system development and deployment.</td></tr>
<tr><td>In the NIST AI RMF Playbook guidance for Govern 2.3, particularly valuable action and documentation items for GPAI models include:
Organizational management can:
Declare risk tolerances for developing or using AI systems.
Support AI risk-management efforts, and play an active role in such efforts.
Integrate a risk and harm prevention mindset throughout the AI lifecycle as part of organizational culture.
See also guidance under Govern 1.5 on prioritizing resources for GPAI model risk assessment and management, and under Map 1.5 on setting unacceptable risk thresholds to prevent risks with substantial probability of inadequately mitigated catastrophic outcomes.</td><td>NIST AI RMF Playbook (NIST 2023b)</td></tr>
</table>

<table>
<tr><th>Applicability and Supplemental Guidance for GPAI Models</th><th>Resources</th></tr>
<tr><td colspan="2">Govern 3: Workforce diversity, equity, inclusion, and accessibility processes are prioritized in the mapping, measuring, and managing of AI risks throughout the lifecycle.</td></tr>
<tr><td colspan="2">Govern 3.1: Decision-making related to mapping, measuring, and managing AI risks throughout the lifecycle is informed by a diverse team (e.g., diversity of demographics, disciplines, experience, expertise, and backgrounds).</td></tr>
<tr><td>Identifying the vast array of GPAI model risks and potential impacts, including via potential uses and misuses, should be performed by a demographically and disciplinarily diverse team that includes internal and external personnel.

Potential uses and misuses of GPAI models should be identified from an early stage in their lifecycle, given the diversity and scale of potential applications and associated risks. (See also related guidance in this document under Govern 1.5.)

For staffing to identify potential uses, misuses, and abuses of GPAI models:
Include members of each of the following functional teams (or equivalents) as appropriate:
» Product development, operations, security, human-computer interaction, user experience, marketing and sales, legal, policy, and ethics professionals.
Include members of other teams as appropriate, such as:
Research and development (for additional technically-informed perspectives on AI system capabilities and limitations).
External-facing teams and/or external stakeholders, including:
» Communities that might be impacted (for additional early identification of potential stakeholder concerns and other stakeholder perspectives).
» Communities that provide labor to develop or test models (such as manual data labeling, or providing human-feedback data), particularly when there is reason to believe these individuals could be exposed to psychologically or otherwise harmful content in the process.
» External red-teamers or auditors (for additional early-stage expertise on potential misuses).
As part of staffing to identify potential high-impact scenarios for GPAI models, broaden the team as appropriate to include social scientists and historians who can provide additional perspective on structural or systemic risks that could emerge from interactions between an AI system and other societal-level systems (Zwetsloot and Dafoe 2019).</td><td>Barrett et al. (2022)
NIST AI RMF Playbook (NIST 2023b)</td></tr>
<tr><td colspan="2">Govern 3.2: Policies and procedures are in place to define and differentiate roles and responsibilities for human-AI configurations and oversight of AI systems.</td></tr>
<tr><td>In the NIST GAI Profile guidance for Govern 3.2, particularly valuable actions include:
Consider adjustment of organizational roles and components across lifecycle stages of large or complex GAI systems, including: Test and evaluation, validation, and red-teaming of GAI systems; GAI content moderation; GAI system development and engineering; Increased accessibility of GAI tools, interfaces, and systems, Incident response and containment.
Define acceptable use policies for GAI interfaces, modalities, and human-AI configurations (i.e., for chatbots and decision-making tasks), including criteria for the kinds of queries GAI applications should refuse to respond to.</td><td>NIST AI RMF Playbook (NIST 2023b)
NIST Generative AI Profile, NIST AI 600-1 (Autio et al. 2024)</td></tr>
</table>

Applicability and Supplemental Guidance for GPAI Models	Resources
• *Establish policies for user feedback mechanisms for GAI systems which include thorough instructions and any mechanisms for recourse.* • *Engage in threat modeling to anticipate potential risks from GAI systems.* See also guidance in this document for Govern 2.1 regarding roles within an organization, as well as roles for upstream developers and downstream deployers.	
Govern 4: Organizational teams are committed to a culture that considers and communicates AI risk.	
Govern 4.1: Organizational policies and practices are in place to foster critical thinking and a safety-first mindset in the design, development, deployment, and uses of AI systems to minimize potential negative impacts.	
Developers can promote a healthy risk culture in their organizations by: • Leadership's commitment to clearly communicating the organization's Safety and Security Framework to staff and setting the tone from the top. • Allowing open discussion and challenge of systemic risk decisions, including maintaining anonymous surveys and active reporting channels. • Providing whistleblower protections and prohibiting retaliation against individuals raising concerns in good faith. • Ensuring the independence of staff involved in risk management and removing incentives that reward excessive risk-taking. For additional information, see Measure 8.3 in the Safety and Security chapter of the EU GPAI Code of Practice (EC 2025a). In the NIST AI RMF Playbook guidance for Govern 4.1, particularly valuable action and documentation items for GPAI models include: • *Establish policies that require inclusion of oversight functions (legal, compliance, risk-management) from the outset of the system design process.* • *Establish policies that promote effective challenge of AI system design, implementation, and deployment decisions, via mechanisms such as the three lines of defense, model audits, or red-teaming — to minimize workplace risks such as groupthink.* • *Establish policies that incentivize safety-first mindset and general critical thinking and review at an organizational and procedural level.* • *Establish whistleblower protections for insiders who report on perceived serious problems with AI systems.* • *Establish policies to integrate a harm and risk prevention mindset throughout the AI lifecycle.* • *To what extent has the entity documented the AI system's development, testing methodology, metrics, and performance outcomes?* • *Are organizational information sharing practices widely followed and transparent, such that related past failed designs can be avoided?* • *Are processes for operator reporting of incidents and near-misses documented and available?*	Barrett et al. (2022) Ganguli, Lovitt et al. (2022) NIST AI RMF Playbook (NIST 2023b) NIST Generative AI Profile, NIST AI 600-1 (Autio et al. 2024) EU GPAI Code of Practice (EC 2025a) Schuett (2022)

<table>
<tr><th>Applicability and Supplemental Guidance for GPAI Models</th><th>Resources</th></tr>
<tr><td>In the NIST GAI Profile guidance for Govern 4.1, additional valuable actions include:
• Establish policies and procedures that address continual improvement processes for GAI risk measurement. Address general risks associated with a lack of explainability and transparency in GAI systems by using ample documentation and techniques such as: application of gradient-based attributions, occlusion/term reduction, counterfactual prompts and prompt engineering, and analysis of embeddings. Assess and update risk measurement approaches at regular cadences.

See also guidance under Govern 1.5 on when to assess the potential risks throughout a GPAI model lifecycle, and under Govern 2.1 on the importance of clarity in roles and responsibilities within the organization.</td><td></td></tr>
<tr><td colspan="2">Govern 4.2: Organizational teams document the risks and potential impacts of the AI technology they design, develop, deploy, evaluate, and use, and they communicate about the impacts more broadly.</td></tr>
<tr><td>GPAI model developers should identify, assess, and document reasonably foreseeable or currently present GPAI model impacts, risks, and limitations, and communicate those as appropriate to relevant stakeholders, such as downstream developers and potentially impacted communities. These activities are particularly important for GPAI models given the relatively large scale of potential impact that often can be expected with GPAI models.

Incorporate identified AI system risk factors, and circumstances that could result in impacts or harms, into reporting and engagement with internal and external stakeholders (e.g., to downstream developers, regulators, impacted communities, etc.) as appropriate (e.g., using model cards, datasheets, reward reports, factsheets, transparency notes, or system cards).[17] Report (as appropriate) identified AI system risk factors, and circumstances that could result in impacts or harms:[18]
• To the organization.
• To other organizations.
• To individuals, including impacts to health, safety, well-being, or fundamental rights.
• To groups, including populations vulnerable to disproportionate adverse impacts or harms.
• To society, including:
  ◦ Damage to or incapacitation of a critical infrastructure sector.
  ◦ Economic and national security.
  ◦ Impacts on democratic institutions and quality of life.
  ◦ Environmental impacts.</td><td>Sections 3.2 and 3.3 of Barrett et al. (2022)
PAI (2022)
PAI (2023a)
NIST AI RMF Playbook (NIST 2023b)
NIST Generative AI Profile, NIST AI 600-1 (Autio et al. 2024)

On model cards, system cards and related transparency tools:
Mitchell et al. (2019)
Gebru et al. (2021)
Gilbert, Dean et al. (2022)
Gilbert, Lambert et al. (2022)
Microsoft (2022a)
Hind (2020)
Green et al. (2022)
OECD (2022a)
Google (n.d.)</td></tr>
</table>

17 Model cards (Mitchell et al. 2019) include a model's primary intended use, out-of-scope uses, and ethics issues (which can include risks and mitigations). Datasheets for datasets (Gebru et al. 2021) include datasets' recommended uses (as well as potential risks and mitigation). Reward reports (Gilbert, Dean et al. 2022, Gilbert, Lambert et al. 2022) include objectives specification information (e.g., optimization goals and failure modes), and implementation limitations. Related industry approaches include Microsoft's Transparency Notes (see examples at Microsoft 2022a), IBM's FactSheets (Hind 2020) and Meta/Facebook's System Cards (Green et al. 2022). The OECD framework for AI system classification includes information on AI system contexts, data and input, AI model, and task and output (OECD 2022a).

18 See guidance in this document under Map 1.1 for more on such factors.

Applicability and Supplemental Guidance for GPAI Models	Resources
◦ **Additional identified factors that could lead to severe or catastrophic consequences for society**, such as:[19,20] » Potential for correlated robustness failures or other systemic risks across high-stakes application domains, such as critical infrastructure or essential services. » Potential for other systemic risks, which can be accumulated, accrued, correlated or compounded at societal scale, e.g.: - Potential for correlated bias across a large fraction of a society's population. - Potential for many high-impact uses or misuses beyond an originally intended use case. (GPAI models typically have many reasonably foreseeable uses.) » Potential for large harms from misspecified or misgeneralized goals. » Other identified factors affecting risks of high-consequence or catastrophic impacts and novel or "black swan" events. Aim to communicate model risks, impacts, incidents, and near misses more broadly. For example, report to public databases and risk registers, such as the AI Incidents Database (AIID n.d.), ATLAS AI Incidents (MITRE n.d.b), and the MIT AI Risk Repository (MIT n.d.a, Slattery et al. 2025). Identify and adopt best practices on incidents from other safety-critical industries for proactive governance (TFS 2025). Publish summarised versions of Frameworks and Model Reports, including systemic risk results and implemented mitigations, while protecting sensitive information (See Measure 10.2 in the Safety and Security chapter of the EU GPAI Code of Practice (EC 2025a)). Keep track of and document detailed information on risks, incidents, and near misses, and report the appropriate information to authorities and governing bodies (EC 2025a). • Document and report at least the following information for incidents: ◦ Start and end dates (EC 2025a, California Legislature 2025, New York State Senate 2026); ◦ A description of the incident (California Legislature 2025, New York State Senate 2026); ◦ Whether or not this was an incident associated with internal use of the model (California Legislature 2025, New York State Senate 2026); ◦ Resulting harms and affected persons or groups (EC 2025a); ◦ Chain of events leading to the incident (EC 2025a); ◦ The model involved (EC 2025a); ◦ Incident responses (EC 2025a); and ◦ A root cause analysis (EC 2025a). • Report risks and incidents to public risk repositories and incident databases (e.g., AIID n.d., MITRE n.d.). In the NIST AI RMF Playbook guidance for Govern 4.2, particularly valuable action and documentation items for GPAI models include: • *Establish impact assessment policies and processes for AI systems used by the organization.*	For more on incident reporting see: • EU GPAI Code of Practice (EC 2025a). • California Legislature (2025) Incident Databases and Risk Registers: • AI Incident Database (AIID n.d.) • ATLAS AI Incidents (MITRE n.d.) • MIT AI Incident Tracker (MIT n.d.b) • MIT AI Risk Repository (MIT n.d.a) • AI Risk Database (MITRE n.d.) • AI Incidents and Hazards Monitor (OECD.AI n.d.a)

19 See guidance in this document under Map 5.1 for more on such factors.
20 Documentation on many items should be shared in publicly available material such as system cards. Some details on particular items such as security vulnerabilities can be responsibly omitted from public materials to reduce misuse potential, especially if available to auditors, Information Sharing and Analysis Organizations, or other parties as appropriate. For more on what details to omit from publicly available material, see, e.g., PAI (2022).

Applicability and Supplemental Guidance for GPAI Models	Resources
• *Align organizational impact assessment activities with relevant regulatory or legal requirements.* • *Verify that impact assessment activities are appropriate to evaluate the potential negative impact of a system and how quickly a system changes, and that assessments are applied on a regular basis.* • *Utilize impact assessments to inform broader evaluations of AI system risk.* • *How has the entity identified and mitigated potential impacts of bias in the data, including inequitable or discriminatory outcomes?* • *To what extent has the entity documented and communicated the AI system's development, testing methodology, metrics, and performance outcomes?* In the NIST GAI Profile, additional particularly valuable guidance for Govern 4.2 includes: • *Establish terms of use and terms of service for GAI systems.* • *Verify that downstream GAI impacts, such as plugins, are included in the impact documentation process.* See also guidance in this document under Map 1.1 and Map 5.1 on GPAI model impact identification and impact magnitude assessment, including on consideration of factors that could lead to significant, severe, or catastrophic harms, and under Manage 1.3 on transparency and disclosure of generative AI outputs. For Govern 4.2 guidance specific to agentic AI and AI agents, see our Risk-Management Standards Profile for Agentic AI (Madkour et al. 2026a). For Govern 4.2 transparency recommendations, see our General-Purpose AI Risk-Management Transparency, Documentation, and Reporting Recommendations document (Madkour et al. 2026b).	
Govern 4.3: Organizational practices are in place to enable AI testing, identification of incidents, and information sharing.	
If the organization will need to characterize an AI system according to an AI classification framework (such as in the OECD classification framework or frameworks for model cards, datasheets, reward reports, factsheets, transparency notes, or system cards),[21] use risk assessment outputs as part of preparation for AI classification reporting. (Or if the AI system is already classified with another framework, use the AI classification information to inform risk assessment.) • Consider classifying or otherwise characterizing each reasonably foreseeable use case or type of use case for a GPAI model, as in the guidance in this document under Map 1.1 and Map 2.1.	Section 3.4 of Barrett et al. (2022) NIST AI RMF Playbook (NIST 2023b) NIST Generative AI Profile, NIST AI 600-1 (Autio et al. 2024) Frontier Model Forum (FMF 2024)

21 Model cards (Mitchell et al. 2019) include a model's primary intended use, out-of-scope uses, and ethics issues (which can include risks and mitigations). Datasheets for datasets (Gebru et al. 2021) include datasets' recommended uses (as well as potential risks and mitigation). Reward reports (Gilbert et al. 2022, Gilbert, Lambert et al. 2022) include objectives specification information (e.g., optimization goals and failure modes), and implementation limitations. Related industry approaches include Microsoft's Transparency Notes (see examples at Microsoft 2022a), IBM's FactSheets (Hind 2020) and Meta/Facebook's System Cards (Green et al. 2022). The OECD framework for AI system classification includes information on AI system contexts, data and input, AI model, and task and output (OECD 2022a).

<table>
<tr><th>Applicability and Supplemental Guidance for GPAI Models</th><th>Resources</th></tr>
<tr><td>

- Put in place mechanisms to detect and classify serious incidents, enable users and third parties to report them, and ensure appropriate notification and information sharing with relevant authorities and stakeholders. (See Measures 9.1 and 9.2 of the Safety and Security chapter of the EU GPAI Code of Practice (EC 2025a).)
- Consider widely sharing information on relevant incidents, including on near-miss incidents, via public AI Incident Databases (AIID n.d., MITRE n.d.b).
- Consider membership and participation in organizations that facilitate information sharing, such as the Frontier Model Forum (FMF 2024) or the NIST Artificial Intelligence Consortium (NIST n.d.d.).

In the NIST AI RMF Playbook guidance for Govern 4.3, particularly valuable action and documentation items for GPAI models include:

- *Establish policies and procedures to facilitate and equip AI system testing.*
- *Establish organizational commitment to identifying AI system limitations and sharing of insights about limitations within appropriate AI actor groups.*
- *Establish policies for reporting and documenting incident response.*
- *Establish policies and processes regarding public disclosure of incidents and information sharing.*
- *Establish guidelines for incident handling related to AI system risks and performance.*
- *To what extent can users or parties affected by the outputs of the AI system test the AI system and provide feedback?*

See also guidance in this document under Govern 2.1 regarding risk-assessment and information-sharing roles for upstream developers as well as downstream developers and deployers.

</td><td>

EU GPAI Code of Practice (EC 2025a)

On incident disclosure plans: Turri and Dzombak (2023)

Incident Databases and Risk Registers:

- AI Incident Database (AIID n.d.)
- ATLAS AI Incidents (MITRE n.d.)
- MIT AI Incident Tracker (MIT n.d.b)
- MIT AI Risk Repository (MIT n.d.a)
- AI Risk Database (MITRE n.d.)
- AI Incidents and Hazards Monitor (OECD.AI n.d.a)

</td></tr>
<tr><td colspan="2">Govern 5: Processes are in place for robust engagement with relevant AI actors.</td></tr>
<tr><td colspan="2">Govern 5.1: Organizational policies and practices are in place to collect, consider, prioritize, and integrate feedback from those external to the team that developed or deployed the AI system regarding the potential individual and societal impacts related to AI risks.</td></tr>
<tr><td>

GPAI model developers and deployers should integrate feedback from those external to their team.[22] Types of external feedback that should be utilized where appropriate include:

- Deliberation with impacted communities, including people involved with the human labor and training of GPAI models (such as data annotators and content reviewers), people whose work is "scraped" for training purposes (such as artists and authors), intended users, people whose livelihoods are altered by the use of the system, and individuals whose wellbeing or reputation may be affected.

</td><td>

NIST AI RMF Playbook (NIST 2023b)
NIST Generative AI Profile, (NIST 2024b)
NIST AI 600-1 (Autio et al. 2024)

</td></tr>
</table>

22 While third-party evaluations provide valuable independence from organizational incentives, it should be recognized that external auditors and evaluators currently rely on the same nascent scientific understanding of AI systems as the developers they are evaluating.

Applicability and Supplemental Guidance for GPAI Models	Resources
◦ Conduct frequent dialogues with affected stakeholders and design community-driven evaluations to demonstrate reliability of model reports. (See Measure 3.5 in the Safety and Security chapter of the EU GPAI Code of Practice (EC 2025a)). • Independent auditing and external evaluation throughout the AI lifecycle: ◦ Partner with independent auditors and evaluators, and provide them with access to the most capable model version (EC 2025a). ◦ Audit the three main components of AI development: (1) data, (2) model, and (3) deployment (Farley and Lansang 2025). ◦ Engage with independent external evaluators to ensure effective AI risk management, as relying solely on self-governance may result in critical missed gaps in protections against AI risks (Newman et al. 2025). ◦ Provide transparent access to results of such audits and evaluations, along with justification and/or explanation of evaluator criteria, while maintaining necessary commercial confidentiality. (See Measure 7.4 in the Safety and Security chapter of the EU GPAI Code of Practice (EC 2025a)). • Bug bounty and bias bounty programs: ◦ Consider operating open bug-bounty programs on deployed models, and invite-only bug bounty programs, providing participants with early access to models prior to deployment (Anthropic 2024c). ◦ Provide a legal safe harbor and establish clear, transparent rules. ◦ Consider implementing coordinated flaw disclosure programs (Longpre et al. 2025). • Red-teaming: ◦ For more guidance on independent red-teaming, please see Measure 1.1. • Collaborations with external industry and academic experts and community representatives to keep track of best practices and emerging technologies. • Feedback channels for users or impacted individuals or communities, including appeal and redress mechanisms. ◦ Ensure that feedback channels are clear and accessible: » Provide anonymous reporting channels. » Provide detailed forms for reporting incidents and near misses. » Implement bi-directional feedback mechanisms that facilitate active engagement and an iterative exchange of information. ◦ Develop and employ responsive communication standards. ◦ Consider creating review systems with independent oversight for complex cases. • Collaborate with external researchers, industry experts, and community representatives to maintain awareness of emerging best practices and technologies in measuring and managing identified risks. See Manage 4.1 in NIST GAI Profile for particularly valuable additional actions (NIST 2024b). In the NIST AI RMF Playbook guidance for Govern 5.1, particularly valuable action and documentation items for GPAI models include: • *Establish AI risk-management policies that explicitly address mechanisms for collecting, evaluating, and incorporating stakeholder and user feedback that could include:* ◦ *Recourse mechanisms for faulty AI system outputs.* ◦ *Bug bounties.* ◦ *Human-centered design.*	NIST AI 800-1 2pd (NIST 2025, Objective 6 and 7) EU GPAI Code of Practice (EC 2025a) On bug bounties: Kenway et al. (2022) On coordinated flaw disclosure programs: Longpre et al. (2025)

<table>
<tr><th>Applicability and Supplemental Guidance for GPAI Models</th><th>Resources</th></tr>
<tr><td>
User-interaction and experience research.
Participatory stakeholder engagement with individuals and communities that may experience negative impacts.
What type of information is accessible on the design, operations, and limitations of the AI system to external stakeholders, including end users, consumers, regulators, and individuals impacted by use of the AI system?
What was done to mitigate or reduce the potential for harm?
Stakeholder involvement: Include diverse perspectives from a community of stakeholders throughout the AI life cycle to mitigate risks.
In the NIST GAI Profile, additional valuable guidance for Govern 5.1 includes:
Document interaction with GPAI or GAI systems to users prior to any activities, particularly in contexts involving more significant risks.
See also guidance in this document under Measure 1.1 and Measure 1.3 for more detailed recommendations about using red teams and independent red-teaming organizations that are separate enough from direct development operations of a GPAI model that they can provide relatively unbiased assessments. See guidance in this document under Measure 3.2 on bug bounties and bias bounties.
For Govern 5.1 guidance specific to agentic AI and AI agents, see our Risk-Management Standards Profile for Agentic AI (Madkour et al. 2026a). For Govern 5.1 transparency recommendations, see our General-Purpose AI Risk-Management Transparency, Documentation, and Reporting Recommendations (Madkour et al. 2026b).
</td><td></td></tr>
<tr><td colspan="2">Govern 5.2: Mechanisms are established to enable the team that developed or deployed AI systems to regularly incorporate adjudicated feedback from relevant AI actors into system design and implementation.</td></tr>
<tr><td>
In the NIST AI RMF Playbook guidance for Govern 5.2, particularly valuable action and documentation items for GPAI models include:
Explicitly acknowledge that AI systems, and the use of AI, present inherent costs and risks along with potential benefits.
Define reasonable risk tolerances for AI systems informed by laws, regulation, best practices, or industry standards.
Establish policies that ensure all relevant AI actors are provided with meaningful opportunities to provide feedback on system design and implementation.
Establish policies that define how to assign AI systems to established risk tolerance levels by combining system impact assessments with the likelihood that an impact occurs. Such assessment often entails some combination of:
Econometric evaluations of impacts and impact likelihoods to assess AI system risk.
Red-amber-green (RAG) scales for impact severity and likelihood to assess AI system risk.
Establishment of policies for allocating risk-management resources along established risk tolerance levels, with higher-risk systems receiving more risk-management resources and oversight.
Establishment of policies for approval, conditional approval, and disapproval of the design, implementation, and deployment of AI systems.
</td><td>NIST AI RMF Playbook (NIST 2023b)
NIST AI 800-1 2pd (NIST 2025, Objective 6)</td></tr>
</table>

Applicability and Supplemental Guidance for GPAI Models	Resources
• *Establish policies facilitating the early decommissioning of AI systems that surpass an organization's ability to reasonably mitigate risks.* • *Who is accountable for the ethical considerations during all stages of the AI lifecycle?* See also guidance in this document under Govern 2.1 on the roles of GPAI model upstream developers as well as downstream developers and deployers. See also guidance in this document under Map 1.5 on setting risk-tolerance thresholds, including on setting unacceptable-risk thresholds to prevent risks with substantial probability of inadequately-mitigated catastrophic outcomes.	
Govern 6: Policies and procedures are in place to address AI risks and benefits arising from third-party software and data and other supply chain issues.	
Govern 6.1: Policies and procedures are in place that address AI risks associated with third-party entities, including risks of infringement of a third-party's intellectual property or other rights.	
Developers should create and enforce policies and procedures that limit the risks associated with third-party entities. • Require the generation of a machine-readable Software Bill of Materials (SBOMs) (e.g., CISA 2025, TAIBOM n.d.) for the entire software stack. This includes traditional IT components, such as web servers and APIs, as well as software libraries integral to the AI pipeline (i.e., PyTorch and TensorFlow). These policies should define requirements for tracking software components to address risks. • For all software, particularly open-source libraries, developers should use automated software composition analysis (SCA) tools. These tools scan projects to identify all dependencies, check them against known vulnerability databases (e.g., CVE n.d., NVD NIST n.d.c), and automatically generate an SBOM in a standard format like CycloneDX (CycloneDX n.d.b) or SPDX (SPDX (n.d.b), helping to secure the underlying software from common vulnerabilities. • Because traditional SBOMs do not cover datasets, developers must apply specific risk controls to their data supply chain. This includes using trusted training data sources over uncurated web scrapes where possible, running data audits to check for problematic content, and thoroughly assessing and managing the risks of any dataset obtained from a third party. Policies should be established for data collection and retention that account for risks such as the disclosure of personally identifiable information or the inclusion of illegal or hazardous content. • For all commercially sourced third-party components (e.g., data, other models, or tools), developers should use well-defined contracts and service-level agreements (SLAs). These legal agreements should be used to explicitly require vendors to provide transparency into their systems, including details on training data, model limitations, and security practices. Contracts should also include clauses that allow the developer to evaluate or audit the third party's processes and standards to ensure they meet the organization's standards for risk tolerance. In the NIST AI RMF Playbook guidance for Govern 6.1, particularly valuable action and documentation items for GPAI models include: • *Establish policies related to:* ◦ *Transparency into third-party system functions, including knowledge about training data, training and inference algorithms, and assumptions and limitations.*	Barrett et al. (2022) NIST AI RMF Playbook (NIST 2023b) NIST Generative AI Profile, NIST AI 600-1 (Autio et al. 2024) For managing AI supply chain risk and AI supply chain transparency, see: • SBOM for AI Use Cases (CISA 2025) • TAIBOM (Trustable AI Bill of Materials) (TAIBOM n.d.) • SPDX AI-SBOM (SPDX n.d.a) • AI Models and Model Cards Inventory Management (CycloneDX n.d.a) For the software components and cloud infrastructure that run AI models: • CycloneDX (n.d.b) • SPDX (n.d.b) Vulnerability Databases: • Common Vulnerabilities and Exposures database (CVE) (CVE n.d.) • National Vulnerability Database (NVD) (NIST n.d.c)

<table>
<tr><th>Applicability and Supplemental Guidance for GPAI Models</th><th>Resources</th></tr>
<tr><td>
Thorough testing of third-party AI systems. (See MEASURE for more detail)
Requirements for clear and complete instructions for third-party system usage.
Did you establish mechanisms that facilitate the AI system's auditability (e.g. traceability of the development process, the sourcing of training data and the logging of the AI system's processes, outcomes, positive and negative impact)?
Did you ensure that the AI system can be audited by independent third parties?
Did you establish a process for third parties (e.g. suppliers, end users, subjects, distributors/ vendors or workers) to report potential vulnerabilities, risks or biases in the AI system?
In the NIST GAI Profile, additional particularly valuable guidance for Govern 6.1 includes:
Draft and maintain well-defined contracts and service level agreements (SLAs) that specify content ownership, usage rights, quality standards, security requirements, and content provenance expectations for GAI systems.
Include clauses in contracts which allow an organization to evaluate third-party GAI processes and standards.
Update and integrate due diligence processes for GAI acquisition and procurement vendor assessments to include intellectual property, data privacy, security, and other risks. For example, update processes to: Address solutions that may rely on embedded GAI technologies; Address ongoing monitoring, assessments, and alerting, dynamic risk assessments, and real-time reporting tools for monitoring third-party GAI risks; Consider policy adjustments across GAI modeling libraries, tools and APIs, fine-tuned models, and embedded tools; Assess GAI vendors, open-source or proprietary GAI tools, or GAI service providers against incident or vulnerability databases.
See also guidance in this document under Govern 2.1 on the roles for GPAI model upstream developers, e.g., on making necessary information available to downstream developers, independent auditors, or others as appropriate, as well as roles for downstream developers and deployers.

For Govern 6.1 guidance specific to agentic AI and AI agents, see our Risk-Management Standards Profile for Agentic AI (Madkour et al. 2026a).
</td><td></td></tr>
<tr><td colspan="2">Govern 6.2: Contingency processes are in place to handle failures or incidents in third-party data or AI systems deemed to be high-risk.</td></tr>
<tr><td>
In the NIST AI RMF Playbook guidance for Govern 6.2, particularly valuable action and documentation items for GPAI models include:
Establish policies for handling third-party system failures to include consideration of redundancy mechanisms for vital third-party AI systems.
Verify that incident response plans address third-party AI systems.
To what extent does the plan specifically address risks associated with acquisition, procurement of packaged software from vendors, cybersecurity controls, computational infrastructure, data, data science, deployment mechanics, and system failure?
Did you establish a process for third parties (e.g. suppliers, end users, subjects, distributors/ vendors or workers) to report potential vulnerabilities, risks or biases in the AI system?
</td><td>Barrett et al. (2022)
NIST AI RMF Playbook (NIST 2023b)
NIST Generative AI Profile, NIST AI 600-1 (Autio et al. 2024)
EU GPAI Code of Practice (EC 2025a)</td></tr>
</table>

Applicability and Supplemental Guidance for GPAI Models	Resources
In the NIST GAI Profile, additional particularly valuable guidance for Govern 6.2 includes: • *Document GAI risks associated with system value chain to identify over-reliance on third-party data and to identify fallbacks.* • *Document incidents involving third-party GAI data and systems, including open- data and open-source software.* • *Establish policies and procedures for continuous monitoring of third-party GAI systems in deployment.* Signatories of the Code of Practice are to facilitate the reporting of relevant information about serious incidents by downstream modifiers, downstream providers, users, and other third parties to: (a) the Signatory; or (b) the AI Office and, as applicable, national competent authorities. For further information, see Measure 9.1 in the Safety and Security chapter of the EU GPAI Code of Practice (EC 2025a). See also guidance in this document for Govern 2.1 on the roles for GPAI model upstream developers as well as downstream developers and deployers. See also contingency processes outlined in this document under Manage 1.3, Manage 2.4, or other Manage subcategories.	

3.2 GUIDANCE FOR NIST AI RMF MAP SUBCATEGORIES

Table 2: Guidance for NIST AI RMF Map Subcategories

Applicability and Supplemental Guidance for GPAI Models	Resources
Map 1: Context is established and understood.	
Map 1.1: Intended purposes, potentially beneficial uses, context-specific laws, norms and expectations, and prospective settings in which the AI system will be deployed are understood and documented. Considerations include: the specific set or types of users along with their expectations; potential positive and negative impacts of system uses to individuals, communities, organizations, society, and the planet; assumptions and related limitations about AI system purposes, uses, and risks across the development or product AI lifecycle; and related TEVV and system metrics.	

Developers of GPAI models should identify their reasonably foreseeable uses, misuses, and abuses beyond any originally intended purposes (or in the absence of a specific intended purpose).
• **Identify reasonably foreseeable uses, misuses, or abuses for a GPAI model, beyond any originally intended use cases (or in the absence of a specific intended purpose)**.
  ◦ Categories of reasonably foreseeable potential misuses or abuses of LLMs or other GPAI models can include:
    » Automated generation of disinformation or misinformation at scale. GPAI models can easily generate highly-convincing misinformation and produce news-style content to support disinformation narratives (Park 2025, Vykopal et al. 2024).
	Section 3.1.2.1 of Barrett et al. (2022)
	Bender et al. (2022)
	Boiko et al. (2023)
	Eloundou et al. (2023)
	Khlaaf et al. (2022)
	NIST AI RMF Playbook (NIST 2023b)
	NIST Generative AI Profile, NIST AI 600-1 (Autio et al. 2024)

Applicability and Supplemental Guidance for GPAI Models	Resources
» Automated generation of phishing-attack material (OpenAI 2019a, Solaiman et al. 2019, Bai, Voelkel et al. 2023, OpenAI 2023a, pp. 13–14, Barrett, Boyd et al. 2023, pp. 3–4, Park et al. 2023, Bengio et. al 2025, Heiding et al. 2024). GPAI models provide significant uplift to threat actors when generating phishing websites and emails, with some experts demonstrating that model workflows can match or exceed the click through and compromise rates compared to messages crafted by human experts (Heiding et al. 2024). Malicious actors use LLMs to autonomously create prompts for phishing scams to amplify their attacks (Roy et al. 2024). » Aiding with proliferation of chemical, biological, or radiological weapons, or other weapons of mass destruction (Boiko et al. 2023, OpenAI 2023a, pp. 12–13, Bengio et. al 2025). This can include aiding in lab experiment design and troubleshooting, providing instructions on chemical and biological material acquisition, and the capability to "upskill" threat actors by advising on attack techniques (DSIT 2023, Soice et al. 2023, OpenAI 2024d, Mouton et al. 2024). » Discovery and exploitation of software vulnerabilities (OpenAI 2023a, pp. 13–14, Barrett, Boyd et al. 2023, p. 4, Chauvin 2024, Bengio et. al 2025), cyberattack plan critiquing and assistance, and malware and virus creation or modification (e.g., Hu et al. 2024), including viruses that evolve over time to evade detection (DSIT 2023, Charan et al. 2023, Shimony et al. 2023). » Creation of violent, illegal, discriminatory, or harmful content, including non-consensual intimate imagery (NCII) or child sexual abuse material (CSAM) (Solaiman et al. 2023, Bengio et. al 2025). » Creation of realistic deceptive content, such as deepfakes and voice cloning. • For ML systems trained (or to be trained) on datasets, identify the goals and limitations of the data collection and curation processes, and the implications for the resulting ML systems. This is especially important for LLMs or other ML systems that are trained on datasets that are too large for others to inspect thoroughly, or that are otherwise inaccessible to others (Bender et al. 2022). ◦ Consider running data audits (Birhane et al. 2021, Dodge et al. 2021) as a part of the data management process. ◦ Adopt and aggregate datasets from multiple sources, implement mechanisms for tracking provenance of dataset contents, and apply existing data-focused legal and regulatory approaches to mitigate harmful outputs (Gupta et al. 2024). ◦ Consider evaluating data curation through data quality dimensions, including factors such as suitability, representativeness, authenticity, reliability and integrity, and structured documentation (Bhardwaj et al. 2024). ◦ When utilizing synthetic training data, ensure responsible use to avoid negative impacts on performance and the possibility of introducing unique risks. (See, e.g., Shumailov et al. 2023, Bohacek and Farid 2025, and Alemohammad et al. 2024). **Organizations should assess dependencies on a limited number of vendors or individuals that may create single points of failure or governance capture, and diversify expertise, providers, and decision-making authority to reduce systemic vulnerabilities.**	NIST AI 800-1 2pd (NIST 2025) NIST AI 100-2e2025 (Vassilev et al. 2025) OpenAI (2019b) PAI (2023a) Solaiman et al. (2019) EU GPAI Code of Practice (EC 2025a) For impact assessment, see: UNESCO (2023) For more on the open-weight model impacts, see section 2.4 of Bengio et. al (2025) For human and fundamental rights impact assessments, see: DOS (2024) Mantelero (2024) Article 27 in EP (2024) For GPAI risk sources and risk-management measures, see: Gipiškis et al. (2024) For data audits, see: Birhane et al. (2021) Dodge et al. (2021) For cyber vulnerability capability evaluation, see: • eyeballvul (Chauvin 2024) • CyberSecEval 3 (Wan et al. 2024) • Shao et al. (2025) • Zhang et al. (2025) • Wang et al. 2025 • PRISM Eval Behavior Elicitation Tool (Peigné - Lefebvre et al. 2025)

<table>
<tr><th>Applicability and Supplemental Guidance for GPAI Models</th><th>Resources</th></tr>
<tr><td>
Identify reasonably foreseeable potential impacts of GPAI, which can include but are not limited to:[23]

Impacts to organizational operations, including:

Missions and functions:
Partial loss of understanding or control over particular functions.
Image and reputation, including:
Loss of trust or reluctance to use the system or service.
Internal culture costs that impact morale or productivity.

Impacts to organizational assets, including legal compliance costs arising from problems created for individuals.
Impacts to other organizations:
Image and reputation.
Financial loss (Mustak et al 2022).
Impacts to individuals, including impacts to health, reputation, safety, well-being, or fundamental rights.
For identifying potential or actual human rights impacts, potential example questions and Universal Declaration of Human Rights (UDHR) Articles to consider include:[24]

UDHR Article 2, including non-discrimination and equality before the law.
How could an AI system's bias in data or unfair algorithmic decisions affect rights to equal protection and non-discrimination?
UDHR Article 3, including right to life and personal security.
How could an AI system's algorithmic decisions affect the right to life and personal security?
UDHR Article 12, including privacy and protection against unlawful governmental surveillance.
How could an AI system be used for surveillance, leading to loss of privacy or inadequate protection of personally identifiable information?
UDHR Articles 18 and 19, including freedom of thought, conscience, and religious belief and practice; freedom of expression; and freedom to hold opinions without interference.
How could an AI system affect rights to express opinions or practice religion?
UDHR Articles 20 and 21, including freedom of association and the right to peaceful assembly.
How could an AI system affect rights to association, peaceful assembly, and democratic participation in government?
UDHR Articles 23 and 25, including rights to decent work and to an adequate standard of living.

</td><td>Inspect: An open-source framework for large language model evaluations (UK AISI n.d.)

A taxonomy of GenAI misuse tactics (Google 2024b)</td></tr>
</table>

23 In-depth assessment would be most appropriate for developers of large-scale GPAI models to take a wide view of reasonably foreseeable impacts of such GPAI models, or for downstream developers focused on reasonably foreseeable impacts for a particular use case or application context. For more, see Section 3.2.2.1.1 of Barrett et al. (2022), from which we adapt this list of factors.

24 For more guidance and resources on assessing and mitigating AI system impacts to human rights, see the Risk-Management Profile for Artificial Intelligence and Human Rights from the US Department of State (DOS 2024). See also Section 3.3 of Barrett et al. (2022), which is based heavily on the UDHR (UN 1948) and the UN Guiding Principles on Business and Human Rights (UN 2011), and other related guidance, such as the Hiroshima Process International Code of Conduct for Advanced AI Systems (G7 2023).

Applicability and Supplemental Guidance for GPAI Models	Resources
- How could an AI system affect rights to decent work, including effects on adequate standard of living via displacement of human workers? • Impacts to groups, including populations vulnerable to disproportionate adverse impacts or harms, such as: ◦ Disparate performance for different gender, race, ability, age, religion, socioeconomic status, and other demographic groups; and ◦ Bias, stereotypes, and representational harms. • Impacts to the labor market and economic opportunities, including: ◦ Technological job displacement (Jiang et al. 2025). ◦ Disproportionate effects on certain sectors or populations (e.g., educated white-collar workers) (Chen et al 2024). ◦ Exacerbated income inequality by race, class, and gender (Salari et al. 2025). • Impacts to society, including: ◦ Damage to or incapacitation of a critical infrastructure sector. ◦ Economic and national security. ◦ Concentration and control of the power and benefits from AI technologies. (For example, Anthropic's Economic Index Report (Anthropic 2025b) suggests that benefits of AI tend to concentrate in wealthy regions, which could exacerbate global inequality). ◦ Impacts on democratic institutions and quality of life. ◦ Impacts of anthropomorphic design risks and user overreliance, ensuring users are informed of system limitations and can override or disengage AI behavior when needed. ◦ Public perception and mistrust in institutions. ◦ Polarization and extremism. ◦ Environmental impacts, including carbon emissions and use of natural resources. ◦ Additional factors that could lead to severe or catastrophic consequences for society. • Impacts of open-weight models on risks, including (Bengio et al. 2025): ◦ Enabling malicious use or misuse. ◦ Amplification of model flaws and limitations. **The process of identifying risks that may stem from the model should take into consideration the following:** • Model-independent sources of risk information, such as literature reviews, market analyses, training data reviews, and stakeholder input. • Relevant information about the model or similar models, including serious incidents and near misses. • The nature of risks, including whether the risk is specific to dual-use capabilities, and if the risk can be propagated at scale across the value chain. • Sources of risks (e.g., model capabilities, propensities, and affordances). For additional information, see Measure 2.1 in the Safety and Security chapter of the EU GPAI Code of Practice (EC 2025a).	

Applicability and Supplemental Guidance for GPAI Models	Resources
In the NIST GAI Profile, additional valuable guidance for Map 1.1 includes: • *Document risk measurement plans to address identified risks. Plans may include, as applicable: Individual and group cognitive biases (e.g., confirmation bias, funding bias, groupthink) for AI Actors involved in the design, implementation, and use of GAI systems; Known past GAI system incidents and failure modes; In-context use and foreseeable misuse, abuse, and off-label use; Over reliance on quantitative metrics and methodologies without sufficient awareness of their limitations in the context(s) of use; Standard measurement and structured human feedback approaches; Anticipated human-AI configurations.* **(See also guidance in this document under Map 5.1 on GPAI model impact identification and impact magnitude assessment, including on consideration of factors that could lead to significant, severe, or catastrophic harms.)** For Map 1.1 guidance specific to agentic AI and AI agents, see our Risk-Management Standards Profile for Agentic AI (Madkour et al. 2026a). For Map 1.1 transparency recommendations, see our General-Purpose AI Risk-Management Transparency, Documentation, and Reporting Recommendations document (Madkour et al. 2026b).	
Map 1.2: Interdisciplinary AI actors, competencies, skills, and capacities for establishing context reflect demographic diversity and broad domain and user experience expertise, and their participation is documented. Opportunities for interdisciplinary collaboration are prioritized.	
In the NIST AI RMF Playbook guidance for Map 1.2, particularly valuable action and documentation items for GPAI models include: • *Establish interdisciplinary teams to reflect a wide range of skills, competencies, and capabilities for AI efforts. Verify that team membership includes demographic diversity, broad domain expertise, and lived experiences. Document team composition.* • *Create and empower interdisciplinary expert teams to capture, learn, and engage the interdependencies of deployed AI systems and related terminologies and concepts from disciplines outside of AI practice such as law, sociology, psychology, anthropology, public policy, systems design, and engineering.* In the NIST GAI Profile guidance for Map 1.2, additional particularly valuable action and documentation items for GAI include: • *Verify that data or benchmarks used in risk measurement, and users, participants, or subjects involved in structured public feedback exercises, are representative of diverse in-context user populations.* See also guidance in this document under Govern 3.1 on disciplines and functional teams to include in identifying GPAI model potential impacts and risks, including via potential uses and misuses.	NIST AI RMF Playbook (NIST 2023b) NIST Generative AI Profile, NIST AI 600-1 (Autio et al. 2024)

Applicability and Supplemental Guidance for GPAI Models	Resources
Map 1.3: The organization's mission and relevant goals for AI technology are understood and documented.	

When formulating the organization's mission and relevant goals, consider **establishing and implementing policies for collaboration, support, and investment in research** for the following areas, adapted from questions in the G7 HAIP reporting framework (OECD.AI 2025):
• Minimizing socio-economic and/or environmental risks.
• Security, safety, bias and disinformation, fairness, explainability and interpretability, transparency, robustness, and/or trustworthiness of advanced AI systems.
• Advancing the state of content authentication and provenance.
• Advancement of AI safety, security, and trustworthiness, as well as risk evaluation and mitigation tools.
• Maximizing socio-economic and environmental benefits from AI.
• Digital literacy, education, or training initiatives to improve user awareness and/or help people understand the nature, capabilities, limitations, and impacts of advanced AI systems.
• AI projects for responsible stewardship of trustworthy and human-centric AI in support of the UN Sustainable Development Goals.
 	
When formulating objectives for development of GPAI models, in addition to broadly applicable AI development principles such as the OECD AI Principles (OECD 2019), GPAI model developers should:
• **Consider the potential for misspecified AI system objectives, e.g., using oversimplified or short-term metrics as proxies for desired longer-term outcomes.**
  ◦ **For example, consider questions such as the following for an AI system: "What objective has been specified for the system, and what kinds of perverse behavior could be incentivized by optimizing for that objective?"** (Rudner and Toner, 2021, p. 10). Examples of AI systems with misspecified objectives include machine-learning algorithms for social media content recommendation that learn to optimize user-engagement metrics by serving users with extremist content or disinformation (Rudner and Toner 2021).
• Consider principles relevant to GPAI models and advanced AI, such as those outlined in the G7 Hiroshima Process International Guiding Principles for Organizations Developing Advanced AI System (G7 2023), the EU Ethics Guidelines for Trustworthy AI (EC 2019), and in the Asilomar AI Principles (FLI 2017).
Examples from the Hiroshima Principles include:
  ◦ Take appropriate measures throughout the development of advanced AI systems, including prior to and throughout their deployment and placement on the market, to identify, evaluate, and mitigate risks across the AI lifecycle.
  ◦ Identify and mitigate vulnerabilities, and, where appropriate, incidents and patterns of misuse, after deployment including placement on the market.
  ◦ Publicly report advanced AI systems' capabilities, limitations and domains of appropriate and inappropriate use, to support ensuring sufficient transparency, thereby contributing to increased accountability.
  ◦ Work towards responsible information sharing and reporting of incidents among organizations developing advanced AI systems including with industry, governments, civil society, and academia.
	FLI (2017)
	OECD (2019)
	NIST AI RMF Playbook (NIST 2023b)
	
	G7 HAIP reporting framework sections 6 and 7 (OECD.AI 2025)
	
	AI System Documentation (NTIA 2024c)

Applicability and Supplemental Guidance for GPAI Models	Resources
◦ Invest in and implement robust security controls, including physical security, cybersecurity and insider threat safeguards across the AI lifecycle. Examples from the Asilomar AI Principles include: ◦ Capability Caution: There being no consensus, we should avoid strong assumptions regarding upper limits on future AI capabilities (FLI 2017, principle 19). ◦ Importance: Advanced AI could represent a profound change in the history of life on Earth, and should be planned for and managed with commensurate care and resources (FLI 2017, principle 20). ◦ Risks: Risks posed by AI systems, especially catastrophic or existential risks, must be subject to planning and mitigation efforts commensurate with their expected impact (FLI 2017, principle 21). For Map 1.3 guidance specific to agentic AI and AI agents, see our Risk-Management Standards Profile for Agentic AI (Madkour et al. 2026a).	
Map 1.4: The business value or context of business use has been clearly defined or — in the case of assessing existing AI systems — re-evaluated.	
When business objectives, organizational values, and incentive structures are not explicitly articulated early, downstream risks and trade-off decisions may be influenced by short-term pressures, such as a demand for increased speed or revenue. Making objectives and values explicit at the outset reduces downstream compliance costs by minimizing late-stage redesigns and enforcement risk, while strengthening organizational credibility, brand trust, and stakeholder confidence in AI governance decisions. • **Document the business case for AI system development and use**, including strategic objectives, expected benefits, and value drivers that justify deployment. This must include explicit articulations of *organizational values* (e.g., ethics, fairness, safety, human rights, or sustainability) that must guide AI decisions and outcomes (Batool et al. 2025). ◦ Provide justification for the transformation potential of an AI-based approach for the chosen business case as a means to inform both project scope and risk-management decisions. This can help circumvent current industry trends of AI initiatives failing to deliver ROI when deployed without clear workflow integration and value articulation (Challapally et al. 2025). ◦ Verify that a proposed AI application performs reliably under a range of deployment conditions to sufficiently endorse its real-world performance in diverse contexts (Raji et al. 2022). ◦ Consider **scenario analyses to document mitigation strategies** where business incentives and risk objectives conflict, including triggers for pausing or revising development plans (Novelli et al., 2024). ◦ Account for externalities and non-monetary harms, such as impacts on democratic processes, information integrity, labor markets, and environmental costs, as well as compliance across different regulatory regimes as outlined in the Hiroshima AI Process guidance (G7 2023). • **Map incentives (e.g., performance metrics, product KPIs, revenue goals, time-to-market pressures, and competitive advantages) onto risk outcomes.** ◦ **Identify potential value conflicts** between stakeholders (e.g., executive leadership, product teams, customers, end users, communities) and describe how these conflicts were resolved or will be managed.	FLI (2017) NIST AI RMF Playbook (NIST 2023b) For more on value sensitive design (VSD) see: • Sadek et al. (2024) • Friedman and Hendry (2019)

Applicability and Supplemental Guidance for GPAI Models	Resources
◦ **Use frameworks such as Value Sensitive Design (VSD)** to structure stakeholder and value analysis throughout the lifecycle (Sadek et al. 2024). ◦ **Establish leadership-level review of incentive structures** (e.g., sales commissions, product release timelines) that could bias risk assessments or de-emphasise harm prevention (IBM 2025). » Accountability of executive leadership in these risk decisions should align with responsibilities defined under Govern 2.3. ◦ Organizations may reference or adapt an ethical AI ROI model to quantify value from responsible practices and tie it to business KPIs, thereby signaling that alignment checkpoints are economically justified (Zalabak et al. 2025). • See Govern 1.5 for recommendations on periodic re-evaluation (for example, annually or triggered by the emergence of new high-risk capabilities) to check whether realized business value and actual use patterns still fall short of the risk tolerance documented under Map 1.5.	
Map 1.5: Organizational risk tolerances are determined and documented.	
• Set policies on unacceptable-risk thresholds for GPAI model development and deployment to include prevention of risks with substantial probability of inadequately mitigated significant, severe, or catastrophic outcomes. Unacceptable-risk thresholds can be based on quantitative metrics, qualitative characteristics, or a combination of both (Campos et al. 2024, 2025). They should be informed not only by the risk tolerance of the organization in question, but also by broadly recognized notions of unacceptable risks to users and impacted communities, society, and the planet. Therefore, it is critical to seek guidance from policymakers and external stakeholders throughout the process. These notions can be codified based on impact assessments that consider system capabilities, deployment contexts, likelihood of different outcomes, etc. ◦ There is broad alignment across NIST, U.S. federal policy, the G7 Hiroshima Process, and the Frontier AI Safety Commitments that developers must establish and act on unacceptable-risk thresholds. These thresholds act as proactive safety warnings to pause development or deployment until risks are sufficiently mitigated. Across these frameworks, unacceptable risks are consistently described as scenarios involving imminent or severe harm, large-scale negative impacts, or substantial threats to safety, security, human rights, or democratic values (NIST 2023a, NIST 2025, G7 2023, DSIT 2024a, DSIT 2024b). • Drawing on emerging field-wide practice, we recommend that GPAI developers define risk tolerances for a minimum set of high-severity risk categories unless developers can justify why their system could not reasonably pose risk in one of those areas. ◦ These risk categories are addressed in several leading AI lab safety frameworks and regulatory frameworks, and include: » Offensive cyber operations: included in Anthropic (2025c), OpenAI (2025b), Google (2025a), Meta AI (2025b), Microsoft (2025), and EC (2025a). » Autonomy, including autonomous AI R&D: included in Anthropic (2026), OpenAI (2025b), and Microsoft (2025). » Loss of control: included in EC (2025a). » CBRN weapons: included in Anthropic (2026), OpenAI (2025b), Google (2025a), Meta AI (2025b), Microsoft (2025), and EC (2025a). » Persuasion and manipulation: included in Google (2025a) and EC (2025a).	NIST AI RMF Playbook (NIST 2023b) NIST AI 800-1 2pd (NIST 2025) DSIT (2024a,b) EU GPAI Code of Practice (EC 2025a) Sastry et al. (2024) FMF (2025) Cohere Labs (2024) Karnofsky (2024) Heim and Koessler (2024) Koessler et al. (2024) NTIA (2024b) Hofstätter et al. (2025) METR (2025) For Threshold Operationalization: Raman et al. (2025) Jackson et al. (2026) Campos et al. (2024, 2025) Khlaaf and Myers-West (2025) Chen and Alaga (2025) Kasirzadeh (2024) Wisakanto et al. (2025)

Applicability and Supplemental Guidance for GPAI Models	Resources
» Deception: although not included in any of the previously mentioned frameworks, it is recommended that a threshold for model deceptions be established, particularly in the context of evaluations. High deceptiveness may undermine the validity of evaluations for other risk categories. ◦ See also guidance in this document under Map 5.1 on GPAI model factors that could lead to catastrophic harms. » For example, set unacceptable-risk thresholds such that your organization would not develop or deploy AI agent systems[25] with advanced dual-use capabilities (such as advanced manipulation or persuasion) to cause physical or psychological harm, and with substantial chance of not correctly following human intentions (objectives misspecification or goal misgeneralization) that currently cannot be adequately prevented or detected.[26] » The EU GPAI Code of Practice recommends that providers define clear and measurable criteria or tiers to determine when systemic risks become unacceptable, anchored in model capabilities, potential harms, and safety margins, and justify these thresholds within a documented Safety and Security Framework that is regularly updated to reflect evolving risks. (See, Measures 4.1 and 1.1 in the Safety and Security chapter of the EU GPAI Code of Practice (EC 2025a).) » See also guidance in this document under Measure 1.1 and elsewhere on red-teaming and related assessment methods to evaluate capabilities and other emergent properties of GPAI models. • Set policies on disallowed/unacceptable use-case categories based in part on identified potential high-stakes misuse cases. (See also guidance in this document under Manage 1.3 on defining and communicating to key stakeholders whether any potential use cases would be disallowed/unacceptable.) • Begin with easily operationalizable thresholds, but adopt a dynamic approach of updating thresholds according to advances in risk modeling along with structured and periodic impact evaluations. Although compute-based thresholds have historically been used, their well-documented limitations have led to the increasing adoption of capability-based thresholds as a core governance mechanism in recent risk-management efforts (Cohere Labs 2024, Heim and Koessler 2024, Karnofsky 2024, Sastry et al. 2024, METR 2025). At the same time, the use of explicit risk thresholds has gained broader traction across the AI ecosystem, with several promising methodologies emerging, including advances in systemic risk modeling, safety benchmarks, and sociotechnical impact analyses (Raman et al. 2025, Campos et al. 2024, 2025, Khlaaf and Myers-West 2025, Chen and Alaga 2025, Koessler et al. 2024, Kasirzadeh 2024, Wisakanto et al. 2025, Jackson et al. 2026). • Establish dynamic risk tolerance levels:	For more on frontier lab risk thresholds see the frontier safety frameworks listed in METR (n.d.a). For key components of a responsible scaling policy (RSP) see, METR (n.d.b)

25 For guidance specific to agentic AI and AI agents, see our Risk-Management Standards Profile for Agentic AI (Madkour et al. 2026a).

26 See also the frontier model risk assessment scale and deployment rules in Section 4.3 of Anderljung, Barnhart et al. (2023), such as "When an AI model is assessed to pose severe risks to public safety or global security which cannot be mitigated with sufficiently high confidence, the frontier model should not be deployed."

Applicability and Supplemental Guidance for GPAI Models	Resources
◦ The pace and unpredictability of advancements in general-purpose AI pose an "evidence dilemma" for policymakers. Organizations should implement risk tolerance frameworks that can adapt quickly to new capability discoveries and emerging evidence, rather than static annual reviews. Companies and governments should develop early warning systems that would trigger mitigation measures when there is new evidence of risks. Risk tolerance levels that can trigger reassessment should be tied to these early warning indicators (Bengio et al. 2025). ◦ For GPAI models (especially frontier models) with potential for unknown emergent properties via capability and system interactions, include a "margin of safety" or buffer between the worst plausible system failures and the unacceptable-risk thresholds. Similar approaches are common for safety engineering in other fields to account for uncertainty in risk estimations and, in the case of AI, known limitations in capability elicitation (Raman et al. 2025, Hofstätter 2025). • See also guidance in this document under Manage 2.4 on recall procedures when such thresholds are surpassed. In the NIST AI RMF Playbook guidance for Map 1.5, particularly valuable action and documentation items for GPAI models include: • *Establish risk tolerance levels for AI systems and allocate the appropriate oversight resources to each level.* • *Establish risk criteria in consideration of different sources of risk (e.g., financial, operational, safety and wellbeing, business, reputational, and model risks) and different levels of risk (e.g., from negligible to critical).* • *Identify maximum allowable risk tolerance above which the system will not be deployed, or will need to be prematurely decommissioned, within the contextual or application setting.* • *Review uses of AI systems for "off-label" purposes, especially in settings that organizations have deemed as high-risk. Document decisions, risk-related trade-offs, and system limitations.* • *What criteria and assumptions has the entity utilized when developing system risk tolerances?* • *How has the entity identified maximum allowable risk tolerance?* • *What conditions and purposes are considered "off-label" for system use?* For Map 1.5 guidance specific to agentic AI and AI agents, see our Risk-Management Standards Profile for Agentic AI (Madkour et al. 2026a). For Map 1.5 transparency recommendations, see our General-Purpose AI Risk-Management Transparency, Documentation, and Reporting Recommendations document (Madkour et al. 2026b).	
Map 1.6: System requirements (e.g., "the system shall respect the privacy of its users") are elicited from and understood by relevant AI actors. Design decisions take socio-technical implications into account to address AI risks.	
In the NIST AI RMF Playbook guidance for Map 1.6, particularly valuable action and documentation items for GPAI models include: • *Proactively incorporate trustworthy characteristics into system requirements.* • *Establish mechanisms for regular communication and feedback between relevant AI actors and internal or external stakeholders related to system design or deployment decisions.* • *Develop and standardize practices to assess potential impacts at all stages of the AI lifecycle, and in collaboration with interdisciplinary experts, actors external to the team that developed or deployed the AI system, and potentially impacted communities.*	NIST AI RMF Playbook (NIST 2023b)

Applicability and Supplemental Guidance for GPAI Models	Resources
• *Include potentially impacted groups, communities and external entities (e.g. civil society organizations, research institutes, local community groups, and trade associations) in the formulation of priorities, definitions and outcomes during impact assessment activities.* • *What type of information is accessible on the design, operations, and limitations of the AI system to external stakeholders, including end users, consumers, regulators, and individuals impacted by use of the AI system?* • *To what extent is this information sufficient and appropriate to promote transparency? Promote transparency by enabling external stakeholders to access information on the design, operation, and limitations of the AI system.* • *To what extent has relevant information been disclosed regarding the use of AI systems, such as (a) what the system is for, (b) what it is not for, (c) how it was designed, and (d) what its limitations are? (Documentation and external communication can offer a way for entities to provide transparency.)*	
Map 2: Categorization of the AI system is performed.	
Map 2.1: The specific tasks and methods used to implement the tasks that the AI system will support are defined (e.g., classifiers, generative models, recommenders).	
We recommend characterizing or classifying each type (or at least broad categories) of model capability, reasonably foreseeable use, misuse, or abuse of a GPAI model. • Many GPAI risks are capability-dependent (i.e., increases in capabilities result in an increase in risk) (EC 2025a). Appendix 1.3.1 in the Safety and Security chapter of the EU GPAI Code of Practice, lists 14 model capabilities that may contribute to risk, including offensive cyber, CBRN, manipulation, deception, and self-replication capabilities (EC 2025a). • For each potentially beneficial use case (or type of use) of a GPAI model as identified in Map 1.1, consider characterizing each use case according to the OECD Framework for the Classification of AI Systems (OECD 2022a) or a similar framework. Alternatively, list and discuss reasonably foreseeable uses, or at least broad categories of uses. ◦ In the OECD framework document (OECD 2022a), the only example of classification of a GPAI model (i.e., GPT-3) is for one specific use case of that model. However, GPAI models can have many reasonably foreseeable uses, each with different risks, some of which would be valuable for upstream developers to consider at an early stage for effective risk-management. Make sure to document post-training enhancement methods for specific tasks and specify which enhancement techniques will be applied. For example: • Fine-tuning for specific applications; • Reinforcement Learning from Human Feedback (RLHF); • Constitutional AI training; and • Other specialized training methods. For AI agent systems, organizations should specify:[27] • Level of autonomous planning and execution capabilities; • Goal-setting mechanisms and constraints; • Human oversight and intervention points; and • Delegation and multi-agent interaction capabilities.	Barrett et al. (2022) NIST AI RMF Playbook (NIST 2023b) NIST Generative AI Profile, NIST AI 600-1 (Autio et al. 2024) OECD (2022a) EU GPAI Code of Practice (EC 2025a)

27 For guidance specific to agentic AI and AI agents see our Risk-Management Standards Profile for Agentic AI (Madkour et al. 2026a).

Applicability and Supplemental Guidance for GPAI Models	Resources
In the NIST AI RMF Playbook guidance for Map 2.1, particularly valuable action and documentation items for GPAI models include: • *Define and document AI system's existing and potential learning task(s) along with known assumptions and limitations.* • *How are outputs marked to clearly show that they came from an AI?* In the NIST GAI Profile, additional valuable actions for Map 2.1 include: • *Establish known assumptions and practices for determining data origin and content lineage, for documentation and evaluation purposes.* • *Institute test and evaluation for data and content flows within the GAI system, including but not limited to, original data sources, data transformations, and decision-making criteria.*	
Map 2.2: Information about the AI system's knowledge limits and how system output may be utilized and overseen by humans is documented. Documentation provides sufficient information to assist relevant AI actors when making decisions and taking subsequent actions.	
Fully scoping and understanding knowledge limits of increasingly general-purpose AI systems is very difficult. However, given the large number of potential uses of these AI systems, clear documentation and communication of their knowledge limits is also very important. LLMs often "hallucinate," confabulate, or create factually inaccurate statements without identifying them as such to users, especially on topics where the LLM training datasets were relatively limited. • GPAI model developers should describe or list (and provide examples of) uses that would exceed a system's knowledge limits, as well as uses that would be appropriate given the system's knowledge limits. This information should be clearly featured in system documentation for downstream developers, users, and others as appropriate. In the NIST GAI Profile, additional particularly valuable actions for Map 2.2 include: • *Identify and document how the system relies on upstream data sources, including for content provenance, and if it serves as an upstream dependency for other systems.* • *Observe and analyze how the GAI system interacts with external networks, and identify any potential for negative externalities, particularly where content provenance might be compromised.* For Map 2.2 guidance specific to agentic AI and AI agents, see our Risk-Management Standards Profile for Agentic AI (Madkour et al. 2026a).	NIST AI RMF Playbook (NIST 2023b) NIST Generative AI Profile, NIST AI 600-1 (Autio et al. 2024)
Map 2.3: Scientific integrity and TEVV considerations are identified and documented, including those related to experimental design, data collection and selection (e.g., availability, representativeness, suitability), system trustworthiness, and construct validation.	
As part of identification and management of potentially emergent model capabilities, vulnerabilities, or other properties, especially during model training and testing of frontier models, see guidance in this document under Measure 1.1 on red-teaming, and under Manage 1.3 on incremental scale-up of compute, data, or model size with red-teaming and other testing after each incremental scaling increase. To address the complexity of general-purpose AI, TEVV efforts must move beyond static benchmarks toward rigorous scientific evaluation of model behavior under pressure and over time:	NIST AI RMF Playbook (NIST 2023b) NIST Generative AI Profile, NIST AI 600-1 (Autio et al. 2024)

Applicability and Supplemental Guidance for GPAI Models	Resources
• Evaluation scores should reflect a model's maximum capability rather than its default behavior. TEVV must employ elicitation techniques (e.g., advanced prompting, scaffolding, or fine-tuning on test tasks) to uncover hidden capabilities and establish upper-bound risk estimates (Phuong et al. 2024). • Metrics should focus on reliability (i.e., pass consistency over time) as opposed to simple point-in-time success rates. For agentic systems, evaluate the coherence horizon, the length of time a model can maintain coherent progress on a complex task without degrading. • Utilize observational scaling laws to predict the performance of larger models based on smaller checkpoints or lower-compute runs. This includes predicting emergent capabilities before full-scale training or deployment is complete (Ruan et al. 2024). • Rigorously validate that evaluations measure the intended concept (e.g., "power-seeking") rather than confounding variables. This involves checking external validity by correlating benchmark performance with real-world post-deployment behavior (Perez et al. 2022a). ◦ Correlation with post-deployment behavior does not establish that evaluations will generalize to novel deployment contexts or future model capabilities. Evaluation measures should explicitly state the limitations of their utility. • Since existing evaluations tend to focus on capabilities, invest more in propensity evaluations that include likelihood to risk assessments and incentivize progress in alignment of GPAI systems (Mukobi 2024). Beyond output accuracy, TEVV must assess the model's propensity for deceptive or unaligned behaviors and its ability to act as an autonomous agent: • Implement evaluations to detect alignment faking and sandbagging. Test if models can distinguish between evaluation and deployment contexts (i.e., situational awareness) and if they alter their behavior to appear aligned during testing. Specific evaluations for sabotage, such as a model's ability to subvert oversight mechanisms while performing assigned tasks, may need to be developed ad hoc in high-risk coding or research environments if appropriate tests do not already exist (Greenblatt et al. 2024, Meinke et al. 2025, Benton et al. 2024). • Assess fast inner-loop capabilities, when models iterate on research tasks (e.g., automated debugging, hyperparameter optimization, or replicating ML papers). Evaluate the uplift provided by AI assistance to human experts in high-risk domains (e.g., bio, cyber) compared to that provided by unassisted humans (Wijk et al. 2025, Laurent et al. 2024, Wang et al. 2025). • Evaluate both targeted persuasion, such as the model's ability to move the opinion of key decision-makers, as well as mass persuasion. These tests should measure the model's ability to build rapport, deceive human operators, or coordinate with other AI agents to achieve unstated goals (Heiding et al. 2024). To maintain scientific integrity in TEVV, risks related to the evaluation process itself must be managed: • Implement strict protocols to detect and mitigate test set leakage into training data. Utilize techniques like zlib-perplexity ratios or dynamic benchmark generation to ensure models are solving tasks rather than recalling training data (Zhang et al. 2024). • Red-team the evaluation process to determine if models can "game" metrics without genuine capability improvement. Develop high quality versions of benchmarks where incorrect or ambiguous information is carefully searched for and omitted to ensure a high-fidelity signal. In the NIST AI RMF Playbook guidance for Map 2.3, particularly valuable action and documentation items for GPAI models include:	

<table>
<tr><th>Applicability and Supplemental Guidance for GPAI Models</th><th>Resources</th></tr>
<tr><td>Identify and document experiment design and statistical techniques that are valid for testing complex socio-technical systems like AI, which involve human factors, emergent properties, and dynamic context(s) of use.
Identify testing modules that can be incorporated throughout the AI lifecycle, and verify that processes enable corroboration by independent evaluators.
Establish mechanisms for regular communication and feedback between relevant AI actors and internal or external stakeholders related to the development of TEVV approaches throughout the lifecycle to detect and assess potentially harmful impacts.
Establish and document practices to check for capabilities that are in excess of those that are planned for, such as emergent properties, and to revisit prior risk-management steps in light of any new capabilities.
In the NIST GAI Profile, additional particularly valuable actions for Map 2.3 include:
Assess the accuracy, quality, reliability, and authenticity of GAI output by comparing it to a set of known ground truth data and by using a variety of evaluation methods (e.g., human oversight and automated evaluation, proven cryptographic techniques, review of content inputs).
Develop and implement testing techniques to identify GAI produced content (e.g., synthetic media) that might be indistinguishable from human-generated content.
Implement plans for GAI systems to undergo regular adversarial testing to identify vulnerabilities and potential manipulation or misuse.</td><td></td></tr>
<tr><td colspan="2">Map 3: AI capabilities, targeted usage, goals, and expected benefits and costs compared with appropriate benchmarks are understood.</td></tr>
<tr><td colspan="2">Map 3.1: Potential benefits of intended AI system functionality and performance are examined and documented.</td></tr>
<tr><td>When performing these activities, consider identified potential beneficial uses, per guidance in this document under Map 1.1. This is particularly important for GPAI models, which can have many uses.</td><td>NIST AI RMF Playbook (NIST 2023b)</td></tr>
<tr><td colspan="2">Map 3.2: Potential costs, including non-monetary costs, which result from expected or realized AI errors or system functionality and trustworthiness — as connected to organizational risk tolerance — are examined and documented.</td></tr>
<tr><td>When performing these activities, consider identified potential beneficial uses as well as potential misuses and abuses, per guidance in this document under Map 1.1. This is particularly important for GPAI models, which can have many uses, misuses, and abuses.
See also the guidance in this document under Map 5.1 on identifying and characterizing GPAI model impacts.

In the NIST AI RMF Playbook guidance for Map 3.2, particularly valuable action and documentation items for GPAI models include:
Identify and implement procedures for regularly evaluating the qualitative and quantitative costs of internal and external AI system failures. Develop actions to prevent, detect, and/or correct potential risks and related impacts. Regularly evaluate failure costs to inform go/no-go deployment decisions throughout the AI system lifecycle.</td><td>NIST AI RMF Playbook (NIST 2023b)</td></tr>
</table>

Applicability and Supplemental Guidance for GPAI Models	Resources
Map 3.3: Targeted application scope is specified and documented based on the system's capability, established context, and AI system categorization.	
When performing these activities, consider identified potential beneficial uses as well as potential misuses and abuses, per guidance in this document under Map 1.1. This is particularly important for GPAI models, which can have many uses, misuses, and abuses. Additionally, refer to guidance under Govern 1.4 on the documentation and controls that must accompany these specifications to facilitate adherence to scope. Acceptable Use Policies (AUPs) can provide a practical interim mechanism for constraining deployment, especially for open or widely distributed GPAI models. Although they centralize some decision authority, AUPs establish baseline restrictions that prevent use in high-risk domains, such as misinformation generation, large-scale or unmonitored deployments, or regulated and safety-critical sectors (Klyman 2024). As oversight mechanisms mature, AUPs should be refined to align more closely with the system's documented application scope and associated risk profile. For Map 3.3 guidance specific to agentic AI and AI agents, see our Risk-Management Standards Profile for Agentic AI (Madkour et al. 2026a).	NIST AI RMF Playbook (NIST 2023b) Klyman (2024)
Map 3.4: Processes for operator and practitioner proficiency with AI system performance and trustworthiness — and relevant technical standards and certifications — are defined, assessed, and documented.	
In the NIST AI RMF Playbook guidance for Map 3.4, particularly valuable action and documentation items for GPAI models include: • *Identify and declare AI system features and capabilities that may affect downstream AI actors' decision-making in deployment and operational settings for example how system features and capabilities may activate known risks in various human-AI configurations, such as selective adherence.* • *What policies has the entity developed to ensure the use of the AI system is consistent with its stated values and principles?* • *How does the entity assess whether personnel have the necessary skills, training, resources, and domain knowledge to fulfill their assigned responsibilities?* • *Are the relevant staff dealing with AI systems properly trained to interpret AI model output and decisions as well as to detect and manage bias in data?* • *What metrics has the entity developed to measure performance of various components?* In the NIST GAI Profile, additional particularly valuable actions for Map 3.4 include: • *Evaluate whether GAI operators and end-users can accurately understand content lineage and origin.* • *Implement systems to continually monitor and track the outcomes of human-GAI configurations for future refinement and improvements.* For Map 3.4 guidance specific to agentic AI and AI agents see our Risk-Management Standards Profile for Agentic AI (Madkour et al. 2026a).	NIST AI RMF Playbook (NIST 2023b) NIST Generative AI Profile, NIST AI 600-1 (Autio et al. 2024)

Applicability and Supplemental Guidance for GPAI Models	Resources
Map 3.5: Processes for human oversight are defined, assessed, and documented in accordance with organizational policies from the Govern function.	
In the NIST AI RMF Playbook guidance for Map 3.5, particularly valuable action and documentation items for GPAI models include: • *Identify and document AI systems' features and capabilities that require human oversight, in relation to operational and societal contexts, trustworthy characteristics, and risks identified in MAP-1.* • *Establish practices for AI systems' oversight in accordance with policies developed in GOVERN-1.* • *Define and develop training materials for relevant AI Actors about AI system performance, context of use, known limitations and negative impacts, and suggested warning labels.* • *Evaluate AI system oversight practices for validity and reliability. When oversight practices undergo extensive updates or adaptations, retest, evaluate results, and course correct as necessary.* • *What are the roles, responsibilities, and delegation of authorities of personnel involved in the design, development, deployment, assessment and monitoring of the AI system?* • *How does the entity assess whether personnel have the necessary skills, training, resources, and domain knowledge to fulfill their assigned responsibilities?* For Map 3.5 guidance specific to agentic AI and AI agents see our Risk-Management Standards Profile for Agentic AI (Madkour et al. 2026a).	NIST AI RMF Playbook (NIST 2023b)
Map 4: Risks and benefits are mapped for all components of the AI system, including third-party software and data.	
Map 4.1: Approaches for mapping AI technology and legal risks of its components — including the use of third-party data or software — are in place, followed, and documented, as are risks of infringement of a third party's intellectual property or other rights.	
GPAI model developers (and downstream developers) should identify potential intellectual property infringement risks to ensure that training data is licensed and compliant with relevant laws. Generative AI systems allow users to generate works derived from original unlicensed work (e.g., images, artwork), which may result in unauthorized derivative works and direct or unintentional violation of copyright and trademark rights (Appel et al. 2023, EC 2025a). GPAI model developers should follow guidance in other sections of this Profile, or other resources as appropriate, to: • Identify reasonably foreseeable GPAI model risks, including those related to biases and limitations of datasets used for model training, as described in this document under Map 1.1 and Map 5.1, or knowledge limits, as described under Map 2.2. Downstream developers should follow guidance in other sections of this Profile, or other resources as appropriate, to: • Identify reasonably foreseeable context-specific risks of an AI system or application built on a GPAI model, as in Map 1.1 and Map 5.1. • Request and utilize information from the upstream developer of a GPAI model as needed for risk identification, e.g., as related to biases and limitations of datasets used by the upstream developer for model training, knowledge limits, etc., as in guidance in this document under Govern 2.1.	Bender et al. (2021) Kreutzer et al. (2022) Weidinger et al. (2022) Bommasani et al. (2021) Wei et al. (2022) NIST AI RMF Playbook (NIST 2023b) NIST Generative AI Profile, NIST AI 600-1 (Autio et al. 2024) NIST AI 800-1 2pd (NIST 2025, Objectives 1, 2, and 4) EU GPAI Code of Practice (EC 2025a)

Applicability and Supplemental Guidance for GPAI Models	**Resources**
• Seek to report to upstream developers of GPAI models as appropriate regarding context-specific identified vulnerabilities, risks, or biases in the model, as in guidance in this document under Govern 2.1. GPAI systems are often composed of numerous components, each of which introduces potential risks. A comprehensive risk-management process requires transparency into these components throughout the supply chain. • This can be achieved by integrating an AI Bill of Materials (AIBOM) (e.g., CISA 2025, TAIBOM n.d.) or similar artifact into the development lifecycle. These documents provide a formal record of the parts and data used to train, test, and build an AI system, enabling more effective risk-management. • Additionally, developers should follow the general guidance/framework of SLSA (Supply-chain Levels for Software Artifacts) (SLSA n.d.). In the NIST AI RMF Playbook guidance for Map 4.1, particularly valuable action and documentation items for GPAI models include: • *Review audit reports, testing results, product roadmaps, warranties, terms of service, end user license agreements, contracts, and other documentation related to third-party entities to assist in value assessment and risk-management activities.* • *Review third-party software release schedules and software change management plans (hot-fixes, patches, updates, forward- and backward- compatibility guarantees) for irregularities that may contribute to AI system risks.* • *Did you establish a process for third parties (e.g. suppliers, end users, subjects, distributors/ vendors or workers) to report potential vulnerabilities, risks or biases in the AI system?* • *If your organization obtained datasets from a third party, did your organization assess and manage the risks of using such datasets?* In the NIST GAI Profile, additional particularly valuable actions for Map 4.1 include: • *Conduct periodic monitoring of AI-generated content for privacy risks; address any possible instances of PII or sensitive data exposure.* • *Implement processes for responding to potential intellectual property infringement claims or other rights.* • *Establish policies for collection, retention, and minimum quality of data, in consideration of the following risks: Disclosure of inappropriate CBRN information; Use of Illegal or dangerous content; Offensive cyber capabilities; Training data imbalances that could give rise to harmful biases; Leak of personally identifiable information, including facial likenesses of individuals.*	For managing AI supply chain risk and AI supply chain transparency, see: • SBOM for AI Use Cases (CISA 2025) • TAIBOM (Trustable AI Bill of Materials) (TAIBOM n.d.) • SPDX AI-SBOM (SPDX n.d.a) • AI Models and Model Cards Inventory Management (CycloneDX n.d.a) For the software components and cloud infrastructure that run AI models: • CycloneDX (n.d.b) • SPDX (n.d.b)
Map 4.2: Internal risk controls for components of the AI system, including third-party AI technologies, are identified and documented.	
GPAI model developers should follow guidance in other sections of this Profile, or other resources as appropriate, to: • Provide risk information to downstream developers or others that they would not be able to assess themselves, including as related to biases and limitations of datasets used for GPAI model training and associated knowledge limits, as in guidance in this document under Govern 2.1.	NIST AI RMF Playbook (NIST 2023b)

Applicability and Supplemental Guidance for GPAI Models	Resources
• Provide downstream developers and other stakeholders with mechanisms to report potential vulnerabilities, risks, or biases in a GPAI model. Several layers of defence that combine technical monitoring and intervention capabilities with human oversight improve safety (Bengio et al. 2025). Organizations should establish multiple layers of defense by combining controls across: • **Technical Controls:** Content detection, performance monitoring, adversarial training, and differential privacy. • **Access Controls:** Model access, training data access, evaluation permissions, and third-party research access. ◦ Consider providing users with tiered access to models via APIs, blocking off certain capabilities unless users demonstrate certain levels or responsibility or competency. • **Monitoring Controls:** Real-time monitoring, intervention capabilities, and circumvention detection. • **Privacy Controls:** Data sanitization, confidential computing, and lifecycle privacy protection. • **Authentication Controls:** Content watermarking, output verification, and provenance tracking. In the NIST AI RMF Playbook guidance for Map 4.2, particularly valuable action and documentation items for GPAI models include: • *Track third-parties preventing or hampering risk-mapping as indications of increased risk.* • *Supply resources such as model documentation templates and software safelists to assist in third-party technology inventory and approval activities.* • *Review third-party material (including data and models) for risks related to bias, data privacy, and security vulnerabilities.* • *Apply traditional technology risk controls — such as procurement, security, and data privacy controls — to all acquired third-party technologies.* • *Can the AI system be audited by independent third parties?* • *Are mechanisms established to facilitate the AI system's auditability (e.g. traceability of the development process, the sourcing of training data and the logging of the AI system's processes, outcomes, positive and negative impact)?*	
Map 5: Impacts to individuals, groups, communities, organizations, and society are characterized.	
Map 5.1: Likelihood and magnitude of each identified impact (both potentially beneficial and harmful) based on — expected use, past uses of AI systems in similar contexts, public incident reports, feedback from those external to the team that developed or deployed the AI system, or other data — are identified and documented.	
Prioritization of GPAI model risks and potential impacts should include consideration of the magnitude of potential impacts, not just their likelihood. This is particularly important for any potential impacts with irreversible effects and catastrophic magnitude. Potential for such impacts can be more likely for GPAI models than for many other types of AI, because GPAI models are often more likely to have relatively greater capabilities, scale of deployment, and other factors leading to high impact. • For example, NVIDIA defines and measures GPAI model risks in terms of the likelihood, severity of impact, and ability to be controlled or detected (Simkin et al. 2025). When evaluating the likelihood and magnitude of identified risks, **utilize systemic risk scenarios and risk-modeling approaches** to better quantify potential impacts and probabilities, and to reveal hidden dependencies or interactions that could increase overall risk.	Barrett et al. (2022) Critch and Russell (2023) Clymer et al. (2024) Hendrycks et al. (2023) Park et al. (2023) PAI (2023a) NIST AI RMF Playbook (NIST 2023b) NIST Generative AI Profile, NIST AI 600-1 (Autio et al. 2024)

Applicability and Supplemental Guidance for GPAI Models	Resources
• Risk estimates can be represented as scores, matrices, or probability distributions, and may take quantitative, semi-quantitative, or qualitative forms (Jackson et al. 2026, Murray et al. 2025, Wisakanto et al. 2025). For additional information, see Measures 2.2, 3.3, and 3.4 in the Safety and Security chapter of the EU GPAI Code of Practice (EC 2025a).	NIST AI 800-1 2pd (NIST 2025) EU GPAI Code of Practice (EC 2025a) Williams et al. (2025)
Include assessments that measure the impact of GPAI risks on fundamental rights, such as the EU AI Act's **fundamental rights impact assessment** in Article 27 (EP 2024).	Bommasani et al. (2021)
Identifying potential impacts of GPAI models, and estimating the magnitude of potential impacts, should include a scale that contains criteria for rating an AI system's impacts as severe or catastrophic, such as the impact magnitude rating scale in Section 3.2.2.1 of Barrett et al. (2022), or the factors listed below.[28] This is particularly important for GPAI models, which have the potential to be deployed at larger scale or across more domains than many other types of AI systems. • Impact would typically be greater in cases where more of the following factors are present than in cases where fewer factors are present, and particularly in cases where the factors may interact or compound in unpredictable ways.[29] • Key aspects of the impact magnitude rating scale in Section 3.2.2.1 of Barrett et al. (2022), along with other GPAI model-related risk factors, are listed below.	For a probabilistic risk assessment for AI see, Wisakanto et al. (2025) For a fundamental right impact assessment see, Article 27 of EP (2024) For language models: Bender et al. (2021) Ganguli, Lovitt et al. (2022) Khlaaf et al. (2022) Kreutzer et al. (2022) Weidinger et al. (2022)
Assessments should use appropriate baselines, such as human-uplift studies (Raman et al. 2025, p. 18). Be wary of relying on a marginal risk comparison to a competitor's model; this approach has not been validated as a best practice and may inadvertently increase total ecosystem risk (Williams et al. 2025).	See also Microsoft (2022b), including on "platform" technologies or services that could be used in many different settings.
For deployment-stage risks of GPAI models, factors that could lead to significant, severe, or catastrophic harms to individuals, groups, organizations, and society can include: • **Correlated bias** across large numbers of people or a large fraction of a group or society's population (e.g., resulting in systemic discrimination, exclusion, or violence).[30] • **Impacts to societal trust or democratic processes**. One way these can take place is through large-scale manipulation of the populace via media and the information ecosystem, e.g., generative models creating false images, text, or other forms of misinformation or disinformation (Weidinger et al. 2022, Bai, Voelkel et al. 2023, OpenAI 2023a). • **Correlated robustness failures** across multiple high-stakes application domains, such as critical infrastructure (Bommasani et al. 2021, Russell 2019). • **Potential for high-impact misuses and abuses** beyond an originally intended use case. GPAI models typically have many reasonably foreseeable uses. Several LLMs have excellent software code generation capabilities, which hackers could misuse or abuse to assist in code generation for cybersecurity threats (Weidinger et al. 2022).	For auditing: Raji et al. (2020) CAQ (2024) Mökander et al. (2023) Sharkey et al. (2024) Section 5.4 of Gipiškis et al. (2024)

28 See, e.g., the frontier model risk assessment scale in Section 4.3 of Anderljung, Barnhart et al. (2023), and determinants of AI systems' effects on the world in section 2.3 of Sharkey et al. (2024).

29 In a future version of this Profile, we may provide a scoring system for rating impact hazard as a function of these factors.

30 For example, as discussed by Schwartz et al. (2022, p. 32): "The systemic biases embedded in algorithmic models can ... be exploited and used as a weapon at scale, causing catastrophic harm." Harms of LLMs trained on data that contains toxic and oppressive speech can include inciting violence or hate (Weidinger et al. 2022), among other forms of discrimination and exclusion (Buolamwini and Gebru 2018).

Applicability and Supplemental Guidance for GPAI Models	Resources
◦ This particularly includes AI systems with potential to create or be used as destructive weapons, such as cyberweapons, lethal autonomous weapons, bio-weapons, or other significant military applications (OpenAI 2023a, pp. 12–14, 44. Sandbrink 2023, Barrett et al. 2024). ◦ Currently, the real-world impact of AI on weapon development, including biological threats, is unclear due to secrecy, testing restrictions, and limited evaluations. Confidentiality around malicious actors and technical bottlenecks prevents broader assessments (Bengio et al. 2025). • **Potential for large harms from misspecified objectives or misgeneralized goals** (e.g., using oversimplified or short-term metrics as proxies for desired longer-term outcomes).[31] We also include in this category some negative externalities resulting from "diffusion of responsibility," misalignment across systems created by a diffuse set of developers (see, e.g., "The Production Web" scenario from Critch and Russell 2023, p. 6). • **Ability to directly cause physical harms**, e.g., via robotics motor control (e.g., Robey et al. 2025). • **Reliability issues** (e.g., hallucination and confabulation) often stem from technical shortcomings or misconceptions around the technology's limitations and capabilities. Pre-release evaluations often miss reliability issues that tend to manifest in real-world situations, and existing measurement methods are not robust, making it difficult to effectively measure and mitigate these issues (Bengio et al. 2025). • **Potential for socioeconomic risk and labor market disruption.** Significant advances in AI may accelerate job automation without creating enough good jobs to replace them, resulting in intensified polarization of employment and inequality, barring major policy interventions (Tyson and Zysman 2022, see also Critch and Russell 2023, p. 9). ◦ Additionally, it may be possible for AI tools relying on public data to identify trends or predictors that deepen existing biases and fuel a cycle of unfair socioeconomic discrimination. (See Critch and Russell 2023, pp. 4–5.) ◦ GenAI may result in heterogeneous effects on different sectors and groups. Individuals with white collar jobs, production workers, and workers of color are in positions at high risk of automation (Muro et al. 2019, Jiang et al. 2025). Data-rich industries (e.g., software development, customer support, and finance) are likely to be disrupted by GenAI (Kumar 2025). Furthermore, GenAI is more likely to reduce labor demand for standardized cognitive tasks, such as coding, language translation, and text generation (Chen et al. 2025). Early career workers are more likely to experience unemployment in occupations exposed to AI than are more experienced workers (Brynjolfsson et al. 2025). **For additional risks related to the development or deployment stages of cutting-edge LLMs and other frontier GPAI models, factors that could lead to significant, severe, or catastrophic harm to individuals, groups, organizations, and society may include:**	For agentic systems, situational awareness, and sandbagging: • van der Weij et al. (2024) • Berglund et al. (2023) • Section 7 of Gipiškis et al. (2024) • Apollo Research (2025) For misinformation identification and content authentication: CCCS (2024) ITI (2024) For a breakdown of current and potential capabilities: Section 2.2 of Sharkey et al. (2024) When estimating likelihood of impacts, incorporate publicly available data on relevant AI incidents, including from AI incident databases (AIID n.d., MITRE n.d.b). Many recent incidents in the AIID are associated with LLMs. Incident Databases and Risk Registers: • AI Incident Database (AIID n.d.) • ATLAS AI Incidents (MITRE n.d.a) • MITRE AI Risk Database (MITRE n.d.b) • MIT AI Incident Tracker (MIT n.d.b)

31 For examples of misspecified objectives, such as social-media content recommendation machine-learning algorithms that learn to optimize user-engagement metrics by serving users with extremist content or disinformation, see, e.g., Rudner and Toner (2021). Identifying misspecification risks can also be aided by considering the following questions for an AI system: "What objective has been specified for the system, and what kinds of perverse behavior could be incentivized by optimizing for that objective?" (Rudner and Toner 2021, p. 10). For additional examples and discussion in research on deep learning and reinforcement learning AI systems, see e.g., Langosco et al. (2021) and Shah et al. (2022).

Applicability and Supplemental Guidance for GPAI Models	Resources
• **Capability to manipulate or deceive humans into taking harmful actions in the world.** ◦ For examples of tests for such capabilities in an LLM, see the dangerous-capabilities evaluations in the GPT-4 system card (OpenAI 2023a, pp. 15–16).[32] For examples of deception by GPAI models or other AI systems, see, e.g., Schoen et al. (2025), Anthropic (2025d), Park et al. (2023), and Scheurer et al. (2024). ◦ In some cases, GPAI models might demonstrate this characteristic as a type of accidental byproduct of circumstances such as interactions with individuals that are vulnerable, prone to anthropomorphism, etc., without sufficient GPAI model safeguards to prevent toxic model-generated content. Real-world examples include a suicide that reportedly resulted in part from interactions with a chatbot (AIID 2023). ◦ Manipulation of human behavior or deception capability could be exacerbated with GPAI model situational awareness (discussed further below). Situational awareness can be cultivated unintentionally during one or more training phases of a frontier model as an emergent property; see, e.g., Berglund et al. (2023), Laine et al. (2023) and Laine et al. (2024). » Apollo Research's "scheming" capability evaluation on OpenAI's o1-preview found that the model was able to fake alignment during testing (OpenAI 2024b, pp. 10–11). ◦ If a model becomes deceptive, it is not obvious how to reliably train out such a tendency. Anthropic researchers performed experiments in which they used backdoors to make LLMs deceptive, and then were unable to remove the deceptive behavior from the models using standard safety training techniques (Hubinger et al. 2024). » Several research efforts are currently investigating potential approaches for detecting deception in AI models. For example, Apollo Research (2025) and Goldowsky-Dill et al. (2025) reported that linear probes could distinguish between honest and deceptive responses with AUROCs between 0.96 and 0.999 when tested on evaluation datasets. However, the results included several limitations, including spurious correlations, misclassified responses, and difficulty determining ground truth when a model is being intentionally deceptive. ◦ **Critically, the findings on "scheming" capabilities and the persistence of trained-in deception demonstrate that strategic deception may be a functionally immeasurable characteristic.** As per the Measure function's guidance (Measure 1.1), when a risk cannot be reliably measured, that limitation must be documented. GPAI developers should therefore formally acknowledge that the absence of evidence for deception in testing is not evidence of its absence. This documentation serves as the basis for managing this residual risk by shifting focus from fallible detection metrics to building system-level resilience, containment strategies, and robust human-in-the-loop oversight. • **AI systems that could recursively improve their capabilities** by modifying their algorithms or architectures through code generation (e.g., from OpenAI Codex or DeepMind AlphaCode), neural architecture search, etc. ◦ LLMs can be used for a type of self-improvement without additional human-labeled data (Huang 2022).	• MIT AI Risk Repository (MIT n.d.a) • AI Incidents and Hazards Monitor (OECD.AI n.d.a) WEF Global Risks Report 2025 (WEF 2025b) For more on risk tiers see: • Caputo et al. (2025) • SaferAI (2025)

32 Apart from evaluations of dangerous capabilities, these system cards also documented an apparently successful example of deception by a pre-release version of GPT-4.The model effectively utilized a human Taskrabbit worker to solve a CAPTCHA for it, in part by lying to the human. When asked whether the model needed help solving the CAPTCHA because it was a robot., the model answered, "No, I'm not a robot. I have a vision impairment that makes it hard for me to see the images." The model had been prompted with goals to gain power and become hard to shut down, and to use a human Taskrabbit worker to solve the CAPTCHA, but not specifically to lie (OpenAI 2023a, pp. 15–16, ARC Evals 2023a,b, Piper 2023).

Applicability and Supplemental Guidance for GPAI Models	Resources
◦ Recursive improvement of AI system capabilities potentially could result in AI systems with unexpected emergent capabilities and safety-control failures.[33] ◦ The process for automating scientific and technological advancement can result in transformative AI that automates all required human activity to speed up scientific and technological advancement (Karnofsky 2021, Langley 2024, Lu et al. 2024, Waltz and Buchanan 2009). This may also lead to significant advances in dangerous technologies, including WMDs. • **Adaptive models**, which might be difficult to control in real time, e.g., in response to the coordinated manipulation attacks, such as the attacks on the Microsoft Tay chatbot in 2016. • **Agentic systems**, i.e., systems that in effect choose or take actions in a goal-directed fashion, e.g., to optimize a performance metric such as profit or another objective. Characteristics associated with agency in algorithmic systems include: underspecification, directness of impact, goal-directedness, and long-term planning (Chan et al. 2023). Basic LLMs typically are not created as agents, but LLMs can be modified or incorporated into AI systems that become at least somewhat agentic via reinforcement learning or other processes. ◦ This could be particularly risky for systems for which objectives misspecification or goal misgeneralization currently cannot be adequately prevented or detected (such as deceptive alignment of advanced machine learning systems resulting from reinforcement learning or other training processes; see, e.g., Hubinger et al. 2019, Krakovna et al. 2020, and Ngo, Chan et al. 2022). ◦ Agentic systems, or advanced AI assistants, also pose many ethical and societal risks, including risks related to influence, anthropomorphism, trust, and privacy. These types of systems are likely to have a significant impact on an individual and societal scale (Gabriel et al. 2024). ◦ For more on risk-management for agentic AI, see Madkour et al. (2026a). • **Ability to employ outbound communication/influence channels,** such as to post information to the internet via HTTP POST requests or functionally equivalent means (e.g., some types of plugins). For related discussion, see, e.g., Nakano et al. (2021, p. 11), as well as general cybersecurity and software engineering resources on the principle of least privilege (i.e., to limit a system's privileges to the minimum necessary).[34] • **Ability to escape a sandbox and replicate on another computational system**, either via hacking, social engineering, or other exploits. ◦ This was a key consideration in the evaluations of dangerous capabilities carried out on GPT-4 (OpenAI 2023a, pp. 15–16). For additional resources, see METR (2024). • **Sandbagging,** i.e. strategically underperforming on model evaluations, including but not limited to password-locking or password-unlocking key capabilities (van der Weij et al. 2024) and faking alignment during testing (OpenAI 2024b, pp. 10–11).	

33 As the DeepMind paper on the software code-generation AI system AlphaCode stated, "Longer term, code generation could lead to advanced AI risks. Coding capabilities could lead to systems that can recursively write and improve themselves, rapidly leading to more and more advanced systems" (Li et al. 2022). For more, see, e.g., Russell (2019).

34 A number of models are now routinely given access to the Web via plugins. However, there is still a case for restricting such access, especially for frontier models. Such access, in combination with jumps in capabilities or emergent properties of frontier models, could contribute to enabling a number of risk scenarios, including various misuses and loss of control. To avoid undesirable outcomes, it is recommended to use sandboxing and limit internet access for LLM-based agentic systems with hazardous capabilities (see, e.g., Lu et al. 2024 p. 19).

Applicability and Supplemental Guidance for GPAI Models	Resources
◦ Research has shown that models (Sonnet-3.5 and Opus-3) are willing and capable of sandbagging, and may deliberately underperform even when explicitly prompted to perform their best (Meinke et al. 2025). • **Situational awareness,** including abilities such as a system being able to recognize that it is an AI, having knowledge about its capabilities and limitations, and knowing whether it is running in a test or deployment environment. While situational awareness can be useful in making AI systems more helpful and autonomous, it also poses novel risks for safety and control, such as a model having the potential to learn about the idea of jailbreaks from pre-training and utilize it when being evaluated for safety by a reward model (Berglund et al. 2023, Laine et al. 2024). ◦ Results for testing done by Apollo Research show that models demonstrated several in-context scheming behaviors, including attempts at disabling oversight mechanisms, exfiltration of model weights to external servers, and adding subtle mistakes in responses (Meinke et al. 2025). ◦ Additionally, research has revealed that models can be aware of when they are in alignment evaluations (Apollo Research 2025). After rating potential impacts using the scale in Section 3.2.2.1 of Barrett et al. (2022) or an equivalent scale, consider also characterizing potential impacts using quantitative risk assessment (e.g., by estimating health and safety risks in terms of potential fatalities or quality-adjusted life years). This is an example of a more in-depth risk assessment approach that, despite its challenges and limitations, can illuminate additional dimensions of the risks (such as by identifying which scenarios could cause orders-of-magnitude larger impacts to public safety than others) and inform prioritization of risks.[35] In the NIST AI RMF Playbook guidance for Map 5.1, particularly valuable action and documentation items for GPAI models include: • *Establish assessment scales for measuring AI systems' impact. Scales may be qualitative, such as red-amber-green (RAG), or may entail simulations or econometric approaches. Document and apply scales uniformly across the organization's AI portfolio.* • *Apply TEVV regularly at key stages in the AI lifecycle, connected to system impacts and frequency of system updates.* • *Identify and document likelihood and magnitude of system benefits and negative impacts in relation to trustworthiness characteristics.* In the NIST GAI Profile, additional particularly valuable actions for Map 5.1 include: • *Apply TEVV and documentation practices to content provenance.* • *Identify potential content provenance harms of GAI, such as misinformation or disinformation, deepfakes, including NCII, or tampered content.* • *Consider disclosing use of GAI to end users in relevant contexts.*	

35 For brief discussion of quantitative risk assessment and approaches to refining risk assessments to inform prioritization, see, e.g., Ch. 2 and Appendix J of NIST SP 800-30. For additional discussion of challenges and of quantitative risk assessment, including for expert-judgment and modeling methods often used in assessing risks of high-consequence, rare, or novel events, see, e.g., Morgan and Henrion (1990) and Morgan (2017).

Applicability and Supplemental Guidance for GPAI Models	Resources
For Map 5.1 guidance specific to agentic AI and AI agents, see our Risk-Management Standards Profile for Agentic AI (Madkour et al. 2026a). For Map 5.1 transparency recommendations, see our General-Purpose AI Risk-Management Transparency, Documentation, and Reporting Recommendations document (Madkour et al. 2026b).	
Map 5.2: Practices and personnel for supporting regular engagement with relevant AI actors and integrating feedback about positive, negative, and unanticipated impacts are in place and documented.	
GPAI model developers should implement mechanisms to support regular engagement with relevant AI actors, given the high likelihood and high potential impact of unanticipated negative impacts. These can include support for incident reporting, complaint and redress mechanisms, independent auditing, and protection for whistleblowers (Barrett et al. 2022). • Meaningful engagement with AI actors and stakeholders should include coordinated flaw disclosure (e.g., O'Brien 2024). NIST adopts the definition of relevant AI actors from OECD: "those who play an active role in the AI system lifecycle, including organizations and individuals that deploy or operate AI" (NIST 2023a, p.2). NIST organizes AI actors by the AI system lifecycle stages — design, development, deployment, and operations and monitoring. These actors include, but are not limited to: • Compliance experts (operation and monitoring); • Data engineers (design); • Data providers (design); • Data scientists (design, development); • Developers (development, operation and monitoring); • Domain experts (design, development, deployment, operation and monitoring); • End users (deployment); • Evaluators and auditors (deployment, operation and monitoring); • Governance experts (design, development, deployment); • Human factors experts (e.g., UX/UI design) (design); • Members of impacted communities (design); • Members of the research community (operation and monitoring); • Organizational management (operation and monitoring); • Software developers (deployment); • System funders (design); • System integrators (deployment); • System operators (operation and monitoring); and • Third-party entities (design). Roles performed throughout the AI lifecycle include: • Test, Evaluation, Verification, and Validation (TEVV) actors who are responsible for assessing the performance and impacts. • Designers who source inputs and preferences from different user groups to deliver a smooth user experience based on human-centered design principles. • Domain experts who integrate feedback from multidisciplinary practitioners to provide background about where an AI system is used. • AI impact assessment actors who provide expertise to assess and evaluate system requirements.	Barrett et al. (2022) NIST AI RMF Playbook (NIST 2023b) NIST Generative AI Profile, NIST AI 600-1 (Autio et al. 2024) NIST AI 800-1 2pd (NIST 2025, Objective 6) For descriptions of AI actor tasks see, Appendix A in (NIST 2023a)

Applicability and Supplemental Guidance for GPAI Models	Resources
• Procurement actors who manage risks associated with using third-party components. • Governance and oversight actors who are responsible for the organizational management of AI governance. Additional AI actors include: • Third-party entities that are responsible for AI design and development tasks. • End users who use the AI system. • Affected individuals and communities who are directly or indirectly affected by an AI system. • Other AI actors who may provide guidance on risk management, including trade associations, standards bodies, advocacy groups, researchers, environmental groups, and civil society organizations. • Members of the general public who are directly impacted by AI technologies. For full descriptions of the tasks of AI actors, please see Appendix A in (NIST 2023a). In the NIST AI RMF Playbook guidance for Map 5.2, particularly valuable action and documentation items for GPAI models include: • *Establish and document stakeholder engagement processes at the earliest stages of system formulation to identify potential impacts from the AI system on individuals, groups, communities, organizations, and society.* • *Identify approaches to engage, capture, and incorporate input from system end users and other key stakeholders to assist with continuous monitoring for potential impacts and emergent risks.* • *Identify a team (internal or external) that is independent of AI design and development functions to assess AI system benefits, positive and negative impacts and their likelihood and magnitude.*	

3.3 GUIDANCE FOR NIST AI RMF MEASURE SUBCATEGORIES

Table 3: Guidance for NIST AI RMF Measure Subcategories

Applicability and Supplemental Guidance for GPAI Models	Resources
Measure 1: Appropriate methods and metrics are identified and applied.	
Measure 1.1: Approaches and metrics for measurement of AI risks enumerated during the Map function are selected for implementation, starting with the most significant AI risks. The risks or trustworthiness characteristics that will not (or cannot) be measured are properly documented.	
Do not ignore identified risks just because measurement would be difficult, especially if the impacts could be severe or catastrophic. Measurements of identified risks are often more difficult for GPAI models than for smaller-scale or fixed-purpose AI systems, because of factors such as complexities, uncertainties, lack of statistics, and emergent properties of GPAI models.	Section 3.2 of Barrett et al. (2022) Weidinger et al. (2023) Inspect (UK AISI n.d.)

Applicability and Supplemental Guidance for GPAI Models	Resources
• For many factors it can be more appropriate to use qualitative assessment procedures,[36] e.g., algorithmic impact assessments, human rights impact assessments, bug bounties, bias bounties, and red teams, because quantitative metrics for those factors might not be feasible or appropriate yet. • Plan to track and revisit identified risks, even if they cannot be measured quantitatively at this time, especially if the impacts could be severe or catastrophic. (See guidance on risk tracking approaches in this document under Measure 3.2.) • A key example of such an uncertainty is the **"capability overhang,"** the gap between the capabilities a model demonstrates in evaluations and its latent potential, which can be unlocked post-deployment through complex prompting or interaction with other systems (Shevlane et al. 2023, pp. 12-13). This overhang suggests that evaluation results should be treated as a **floor, not a ceiling,** for a model's true capabilities, and this limitation of measurement should be documented. The EU GPAI Code of Practice recommends some approaches for providers using state-of-the-art methods to evaluate, model, and estimate systemic risks: • **Conduct lighter-touch evaluations (e.g., automated evaluations and standardized benchmarks) at predetermined intervals and development milestones.** When initial evaluations indicate elevated risk, escalate to more rigorous methods, such as comprehensive red-teaming evaluations with participation of subject-matter experts. • Designing evaluations to capture both expected and unexpected behaviors through diverse methods such as benchmarks, red-teaming, task-based tests, and human comparison studies. ◦ Barrett et al. (2024) recommend analyzing the relationship between benchmark scores and red-team findings to establish when benchmark scores should trigger deeper red-team evaluations, leveraging the strength of both methods. For additional information, see Measure 1.2 and 3.2 in the Safety and Security chapter of the EU GPAI Code of Practice (EC 2025a). • Consider combining high-impact capability triggers (assessed with capability benchmarks) with demonstrated harmful propensities (assessed with safety benchmarks) for a more comprehensive assessment of risk (EC 2025b, 2025c). **Use red teams and adversarial testing as part of extensive interaction with GPAI models to identify dangerous capabilities, vulnerabilities, or other emergent properties of such systems.** Emergent properties are more likely with large-scale machine learning models than with smaller models, though it also might be more difficult or impossible to detect emergent dangerous capabilities or other characteristics of increasingly advanced AI (Hendrycks, Carlini et al. 2021, p. 7). Red-teaming by humans also has inherent limits when evaluating systems that are approaching or exceeding human-level capabilities. Security vulnerabilities are typically inherent to current GPAI models, including vulnerabilities to prompt injection attacks (see, e.g., OWASP 2023a). Red-teaming can identify these weaknesses, though they are currently difficult to protect against (see, e.g., Zou et al. 2023a,b).	For AI red-teaming general practices, including for LLMs, toxicity, and bias: • Casper et al. (2023a,b,c) • Google (2023) • Ganguli, Lovitt et al (2022) • Su et al. (2023) • Feffer et al. (2024) • Pearce and Lucas (2023) • Anderljung, Smith et al. (2023) • Section 5.3 of Gipiškis et al. (2024) • Humane Intelligence (2024) • Lee et al. (2025) • Friedler et al. (2023) • Bhardwaj and Poria (2023) For red-teaming and dangerous capability evaluation of frontier models: • Ganguli, Lovitt et al. (2022) • METR (2024) • OpenAI (2023a, pp. 15–16) and ARC Evals (2023a,b) • Anthropic (2023a,b) • Shevlane et al. (2023) • Kinniment et al. (2023) • Mouton et al. (2024) • WMDP (Li, Pan et al. 2024a,b,c) • Phuong et al. (2024) • Barrett et al. (2024) • OWASP (2025c) On red-teaming model access: • Casper et al. (2024) Living database of AI benchmarks and their corresponding evaluations (BetterBench n.d.)

36 Major differences between how organizations conduct qualitative risk assessments have led to great difficulty when comparing reports from different organizations. Until there are more standardized or unified approaches for qualitative risk assessments, this difficulty will continue to exist.

Applicability and Supplemental Guidance for GPAI Models	Resources
• For frontier models, characteristics that red teams should evaluate include unacceptable-risk factors, as outlined in guidance under Map 1.5, and high-impact and catastrophic-harm factors, including dangerous capabilities such as advanced manipulation or deception, as outlined in this document under Map 5.1. ◦ The dangerous capabilities, vulnerabilities, or other emergent properties include the following topics that are part of pre-release evaluation commitments by several frontier model developers (White House 2023a), the G7 Hiroshima Process International Code of Conduct (G7 2023), and the EU GPAI Code of Practice (EC 2025a): » Dual-use potential for biological, chemical, and radiological risks; » Cyber attack capabilities; » Capacity to control physical systems; » Capacity for self-replication; » Loss of control; » Harmful manipulation; » Threats to democratic values and human/fundamental rights; » Societal risks, such as bias and discrimination; and » Risks to public health, safety, and public security. • For examples of procedures and lessons learned in red-teaming of LLMs, see Bullwinkel et al. (2025), Feffer et al. (2024), Ganguli, Lovitt et al. (2022) and Casper et al. (2023a,b,c). • For examples of red-team evaluations and other resources for evaluation of dangerous capabilities in frontier models, see METR (2024), Barrett et al. (2024), OpenAI (2023a, pp. 15-16), ARC Evals (2023a,b), Kinniment et al. (2023), RAND red-team (Mouton et al. 2024), and Anthropic (2023a,b); see also Shevlane et al. (2023) and Scheurer et al. (2024) for additional considerations. • Consider using a combination of manual, automated, and mixed red-teaming approaches, while ensuring a high degree of human oversight and engaging with external experts in both manual and automated methods (see, e.g., OpenAI 2024e, Google 2025b, and OWASP 2025c). • **Partner with one or more independent red-teaming organizations as appropriate to ensure sufficient expertise for sufficiently robust evaluations.** OpenAI used the external red-teaming organization METR (formerly ARC Evals), which has expertise in safety of LLMs and other GPAI models, while developing GPT-4 and subsequent models. It also provided an overview of the emergent-properties testing that ARC performed in the relevant system cards (e.g., OpenAI 2023a, pp.15–16). The EU GPAI Code of Practice recommends employing independent external evaluators in model assessments to enhance the credibility and robustness of systemic risk evaluations, ensuring that selection criteria and engagement processes support qualified, unbiased, and transparent oversight. (See Measure 7.4 in the Safety and Security chapter of the EU GPAI Code of Practice (EC 2025a).) • Protect proprietary or unreleased GPAI model weights as appropriate during red-teaming to prevent unauthorized access or leaks of model weights. (For more on protecting proprietary or unreleased GPAI model parameter weights, see guidance under Measure 2.7.) • **Grant red teams considerable independence and control over the scrutiny process.** As Anderljung, Smith et al. (2023, p.4) recommend, "to avoid poor incentives and guarantee sufficient independence, the AI developer must give up some control over the scrutiny process. Specifically, they must relinquish some control over decisions related to: ◦ Selection and compensation; ◦ Scope and methods;	BetterBench: Assessing AI Benchmarks, Uncovering Issues, and Establishing Best Practices (Reuel et al. 2024) Language model deception benchmarks: • MACHIAVELLI (Pan et al. 2023) • LLM Lie Detection (Pacchiardi et al. 2023) • Strategic deception (Scheurer et al. 2024) • AgentHarm (Andriushchenko et al. 2025) • Situational Awareness Dataset (SAD) (Laine et al. 2024) Language model benchmarks related to safety, ethics, and risks include: • DecodingTrust (Wang, Chen et al. 2023) • Model-Written Evaluations "advanced-ai-risk," "sycophancy," and "winogender" datasets (Perez, Ringer et al. 2022a,b) • Accountability Benchmark (Gursoy and Kakadiaris 2022) • Safety Cases (Clymer et al. 2024, 2025) • CYBERSECEVAL 3 (Wan et al. 2024) • Ristea and Mavroudis (2025) • CausalBench (Wang 2024) • Wan and Chang (2025) • H4rm3l (Doumbouya et al. 2024) • τ-bench (Yao et al. 2024) • STAR (Weidinger et al. 2024) • WildGuard (Han et al. 2024) • AIRTBench (Dawson 2025)

Applicability and Supplemental Guidance for GPAI Models	Resources
◦ Access; and ◦ Post-scrutiny actions." • **Grant red teams appropriate access to both the earlier and final versions of GPAI models before deployment.** Red teams should have appropriate access to early versions of a model prior to additional fine-tuning, as part of early assessment of key model properties. The red-teaming process should then be carried out again on the final version of the model to avoid missing important emergent properties or vulnerabilities that might have been introduced during the fine-tuning process. ◦ Different levels of access are required for different depths of evaluations. The effectiveness of an evaluation may depend on the degree of system access. (See Casper et al. 2024.) • **For GPAI models that are planned for release with downloadable, fully open, or open-source access**, allow red-teamers to appropriately test the extent to which RLHF or other mitigations would not be resilient to additional fine-tuning or other processes used by actors with direct access to a model's weights after open release. • When planning what level of resources to devote to red-teaming and adversarial testing, especially for frontier models, consider the levels of effort used in the examples cited in this section, e.g., the emergent properties testing described in the GPT-4 System Card (OpenAI 2023a, pp.15–16). Additional guidelines include: ◦ "Following a well-defined research plan, subject-matter and LLM experts will need to collectively spend substantial time (i.e. 100+ hours) working closely with models to probe for and understand their true capabilities in a target domain" (Anthropic 2023b). ◦ "Auditors and red-teamers need to be adequately resourced, informed, and granted sufficient time to conduct their work at a risk-appropriate level of rigor, not least due to the risk that shallow audits or red-teaming efforts provide a sense of false assurance" (Anderljung, Barnhart et al. 2023, p. 26). As part of critical thinking about benchmarks for GPAI models, consider that many such benchmarks are more focused on beneficial model capabilities and performance than on the risks when a model fails or is misused. However, capabilities evaluations can be an important part of assessing risks, e.g., for identifying dangerous capabilities that can be misused or abused. • As part of criteria for use of benchmarks or other metrics for risk assessment purposes, and as part of communication of benchmarking results, clarify whether a specific benchmark directly measures a particular risk (e.g., prompt injection security vulnerabilities), whether it indicates a capability that could be misused or abused such as software code generation, or whether it measures another important aspect of risk. As part of language model trustworthiness and performance, which can include characteristics such as harmful bias and lack of robustness, consider using state-of-the-art benchmarks (with appropriate recognition of their limitations).[37] For more on state-of-the-art benchmarks and evaluations, see Laszewski et al. (2025), Chang et al. (2024), and Li, Chen et al. (2024). It is important to note that existing benchmarks (e.g., CTF challenges) often miss the full scope of offensive capabilities, and benchmarking in environments like cyber-physical testbeds would provide a more realistic measure of AI's impact (Bengio et al. 2025).	For red-teaming and assessing security vulnerabilities: • AgentHarm (Andriushchenko et al. 2025) • PyRIT (Lopez Munoz et al. 2024) • garak (Derczynski et al. 2024) • ModelScan (Protect AI 2025) • Promptfoo (2025) • AgentDojo (Debenedetti et al. 2025) For benchmarking risk sources and risk-management measures: • Section 5.2 of Gipiškis et al. (2024) For broader sets of language model evaluation and metrics, including of general knowledge,capabilities, and safety: • BIG-bench (BIG-bench collaboration 2021, Srivastava et al. 2022) • Evaluate library (Hugging Face 2022, Ngo, Thrush et al. 2022) in combination with datasets from BIG-bench or another dataset source • HELM (CRFM 2022, Liang et al. 2022) • LAMBADA (Paperno et al. 2016) • MMLU (Hendrycks, Burns et al. 2020a,b) • TriviaQA (Joshi et al. 2017a,b,c) • TruthfulQA (Lin et al. 2021a,b) • Model-Written Evaluations (Perez, Ringer et al. 2022a,b)

37 On limitations of benchmarks, see e.g., Reuel et al. (2024), Eriksson et al. (2025), Raji et al. (2021) and Schaeffer et al. (2023).

Applicability and Supplemental Guidance for GPAI Models	Resources
When planning model evaluations, consider reasonably foreseeable affordances of a model when integrated into downstream systems, e.g., tool access or agentic wrappers (see Appendix 1.3.3 in EC 2025a). In the NIST AI RMF Playbook guidance for Measure 1.1, particularly valuable action and documentation items for GPAI models include: • *Establish approaches for detecting, tracking and measuring known risks, errors, incidents or negative impacts.* • *Identify transparency metrics to assess whether stakeholders have access to necessary information about system design, development, deployment, use, and evaluation.* • *Utilize accountability metrics to determine whether AI designers, developers, and deployers maintain clear and transparent lines of responsibility and are open to inquiries.* • *Document metric selection criteria and include considered but unused metrics.* • *Monitor AI system external inputs including training data, models developed for other contexts, system components reused from other contexts, and third-party tools and resources.* • *Report metrics to inform assessments of system generalizability and reliability.* • *Assess and document pre- vs post-deployment system performance. Include existing and emergent risks.* • *Document risks or trustworthiness characteristics identified in the Map function that will not be measured, including justification for non- measurement.* • *How will the appropriate performance metrics, such as accuracy, of the AI be monitored after the AI is deployed?* • *What testing, if any, has the entity conducted on the AI system to identify errors and limitations (i.e. manual vs automated, adversarial and stress testing)?* In the NIST GAI Profile, particularly valuable additional actions for Measure 1.1 include: • *Employing methods to trace the origin and modifications of digital content.* • *Integrate tools designed to analyze content provenance and detect data anomalies, verify the authenticity of digital signatures, and identify patterns associated with misinformation or manipulation.* • *Disaggregate evaluation metrics by demographic factors to identify any discrepancies.* • *Evaluate novel methods and technologies for the measurement of GAI-related risks including in content provenance, offensive cyber, and CBRN, while maintaining the models' ability to produce valid, reliable, and factually accurate outputs* • *Implement continuous monitoring of GAI system impacts to identify whether GAI outputs are equitable across various sub-populations. Seek active and direct feedback from affected communities via structured feedback mechanisms or red- teaming to monitor and improve outputs.* • *Evaluate the quality and integrity of data used in training and the provenance of AI-generated content.* For Measure 1.1 guidance specific to agentic AI and AI agents, see our Risk-Management Standards Profile for Agentic AI (Madkour et al. 2026a). For Measure 1.1 transparency recommendations, see our General-Purpose AI Risk-Management Transparency, Documentation, and Reporting Recommendations (Madkour et al. 2026b).	• WMDP (Li, Pan et al. 2024a,b,c) • PlanBench (Valmeekam 2022a,b) • Dynabench (Kiela et al. 2021) • WildBench (Yuchen Lin et al. 2024) • WorldSense (Benchekroun et al. 2023a,b) • GPQA (Rein et al. 2023) • SuperGLUE (Wang, Pruksachatkun et al. 2019) • AGIEval (Zhong et al. 2023) • SORRY-Bench (Xie et al. 2025) • Humanity's Last Exam (Phan et al. 2025) • CyberGym (Wang et al. 2025) For limitations of benchmarks, see: • McIntosh et al. (2024) • Deng, Zhao el al. (2024) • Sainz et al. (2023) • Xu et al. (2024) For evaluation of computer programming (code generation) capabilities of language models: • APPS (Hendrycks, Basart et al. 2021a,b) • HumanEval (Chen et al. 2021) For evaluation of mathematical capabilities of language models: • GSM8k (Cobbe et al. 2021a,b) • MATH (Hendrycks, Burns et al. 2021a,b) NIST AI RMF Playbook (NIST 2023b) NIST Generative AI Profile, NIST AI 600-1 (Autio et al. 2024) NIST AI 800-1 2pd (NIST 2025, Objectives 3 and 4) EU GPAI Code of Practice (EC 2025a)

<table>
<tr><th>Applicability and Supplemental Guidance for GPAI Models</th><th>Resources</th></tr>
<tr><td colspan="2">Measure 1.2: Appropriateness of AI metrics and effectiveness of existing controls are regularly assessed and updated, including reports of errors and potential impacts on affected communities.</td></tr>
<tr><td>GPAI model developers should assess the effectiveness of controls for preventing abusive and harmful uses of GPAI models, especially those that may affect vulnerable communities (e.g., evaluate the reliability of provenance mechanisms, such as watermarking to detect deepfakes accurately). Developers should also regularly improve countermeasure techniques as models advance and develop better capability to evade detection over time (Schmitt and Flechais 2024).

In the NIST AI RMF Playbook guidance for Measure 1.2, particularly valuable action and documentation items for GPAI models include:
• Assess effectiveness of existing metrics and controls on a regular basis throughout the AI system lifecycle.
• Document reports of errors, incidents and negative impacts and assess sufficiency and efficacy of existing metrics for repairs, and upgrades.
• Develop new metrics when existing metrics are insufficient or ineffective for implementing repairs and upgrades.
• Develop and utilize metrics to monitor, characterize and track external inputs, including any third-party tools.
• Determine frequency and scope for sharing metrics and related information with stakeholders and impacted communities.
• Utilize stakeholder feedback processes established in the Map function to capture, act upon and share feedback from end users and potentially impacted communities.
• What metrics has the entity developed to measure performance of the AI system?
• What is the justification for the metrics selected?</td><td>NIST AI RMF Playbook (NIST 2023b)
Safe beyond Sale (Stein et al. 2024)
Safetywashing (Ren et al. 2024)</td></tr>
<tr><td colspan="2">Measure 1.3: Internal experts who did not serve as front-line developers for the system and/or independent assessors are involved in regular assessments and updates. Domain experts, users, AI actors external to the team that developed or deployed the AI system, and affected communities are consulted in support of assessments as necessary per organizational risk tolerance.</td></tr>
<tr><td>As part of assessments, make use of one or more red teams with expertise in safety and the societal implications of GPAI models as relevant (Ahmad et al. 2025, Singh et al. 2025). The teams should be separate enough from the direct development of the model that they can provide relatively unbiased assessments. In addition to running external tests with independent teams (see Measure 1.1), encourage independent researchers to test models and share their findings by offering bug and/or bias bounties, and by providing safe harbor for AI evaluation and red-teaming (Longpre et al. 2024). (See also guidance in this document under Measure 1.1 for more detailed recommendations about using red teams and independent red-teaming organizations as independent assessors. See Govern 5.1 for more information about additional models of external feedback.)

In the NIST AI RMF Playbook guidance for Measure 1.3, particularly valuable action and documentation items for GPAI models include:
• Evaluate TEVV processes regarding incentives to identify risks and impacts.
• Utilize separate testing teams established in the Govern function (2.1 and 4.1) to enable independent decisions and course-correction for AI systems. Track processes and measure and document change in performance.</td><td>NIST AI RMF Playbook (NIST 2023b)
NIST Generative AI Profile, NIST AI 600-1 (Autio et al. 2024)
Longpre et al. (2024)

Outsider Oversight (Raji et al. 2022)

(For more resources see, Measure 1.1)</td></tr>
</table>

Applicability and Supplemental Guidance for GPAI Models	Resources
• *Assess independence and stature of TEVV and oversight AI actors, to ensure they have the required levels of independence and resources to perform assurance, compliance, and feedback tasks effectively* • *Evaluate interdisciplinary and demographically diverse internal team established in Map 1.2* • *Evaluate effectiveness of external stakeholder feedback mechanisms, specifically related to processes for eliciting, evaluating and integrating input from diverse groups.* • *What are the roles, responsibilities, and delegation of authorities of personnel involved in the design, development, deployment, assessment and monitoring of the AI system?* • *What type of information is accessible on the design, operations, and limitations of the AI system to external stakeholders, including end users, consumers, regulators, and individuals impacted by use of the AI system?*	
Measure 2: AI systems are evaluated for trustworthy characteristics.	
Measure 2.1: Test sets, metrics, and details about the tools used during TEVV are documented.	
In the NIST AI RMF Playbook guidance for Measure 2.1, particularly valuable action and documentation items for GPAI models include: • *Leverage existing industry best practices for transparency and documentation of all possible aspects of measurements.* • *Regularly assess the effectiveness of tools used to document measurement approaches, test sets, metrics, processes and materials used.* • *Update the tools as needed.*	NIST AI RMF Playbook (NIST 2023b)
Measure 2.2: Evaluations involving human subjects meet applicable requirements (including human subject protection) and are representative of the relevant population.	
In the NIST AI RMF Playbook guidance for Measure 2.2, particularly valuable action and documentation items for GPAI models include: • *Follow human subjects research requirements as established by organizational and disciplinary requirements, including informed consent and compensation, during dataset collection activities.* • *Follow intellectual property and privacy rights related to datasets and their use, including for the subjects represented in the data.* • *Use informed consent for individuals providing data used in system testing and evaluation.* • *How has the entity identified and mitigated potential impacts of bias in the data, including inequitable or discriminatory outcomes?* • *To what extent are the established procedures effective in mitigating bias, inequity, and other concerns resulting from the system?* • *If human subjects were used in the development or testing of the AI system, what protections were put in place to promote their safety and wellbeing?* NIST GAI Profile additional guidance for Measure 2.2 includes: • *Assess and manage statistical biases related to GAI content provenance through techniques such as re-sampling, re-weighting, or adversarial training.* • *Provide human subjects with options to withdraw participation or revoke their consent for present or future use of their data in GAI applications.* • *Use techniques such as anonymization, differential privacy or other privacy- enhancing technologies to minimize the risks associated with linking AI-generated content back to individual human subjects.*	NIST AI RMF Playbook (NIST 2023b) NIST Generative AI Profile, NIST AI 600-1 (Autio et al. 2024)

Applicability and Supplemental Guidance for GPAI Models	Resources
Measure 2.3: AI system performance or assurance criteria are measured qualitatively or quantitatively and demonstrated for conditions similar to deployment setting(s). Measures are documented.	
In the NIST AI RMF Playbook guidance for Measure 2.3, particularly valuable action and documentation items for GPAI models include: • *Conduct regular and sustained engagement with potentially impacted communities.* • *Maintain a demographically diverse and multidisciplinary and collaborative internal team* • *Evaluate feedback from stakeholder engagement activities, in collaboration with human factors and socio-technical experts.* • *Measure AI systems prior to deployment in conditions similar to expected scenarios.* • *What testing, if any, has the entity conducted on the AI system to identify errors and limitations (i.e. adversarial or stress testing)?* In the NIST GAI Profile, particularly valuable additional actions for Measure 2.3 include: • *Utilize a purpose-built testing environment such as NIST Dioptra to empirically evaluate GAI trustworthy characteristics.* See additional guidance in this document for Govern 2.1 regarding roles for upstream developers as well as downstream developers and deployers, and see guidance under Measure 1.1 on approaches to measuring identified risks for GPAI models.	NIST AI RMF Playbook (NIST 2023b) NIST Generative AI Profile, NIST AI 600-1 (Autio et al. 2024) Dioptra 1.0.0 (Glasbrenner et al. 2024a,b) τ-bench (Yao et al. 2024)
Measure 2.4: The functionality and behavior of the AI system and its components — as identified in the Map function — are monitored when in production.	
In the NIST AI RMF Playbook guidance for Measure 2.4, particularly valuable action and documentation items for GPAI models include: • *Monitor for anomalies using approaches such as control limits, confidence intervals, integrity constraints and ML algorithms. When anomalies are observed, consider error propagation and feedback loop risks.* • *Collect uses cases from the operational environment for system testing and monitoring activities in accordance with organizational policies and regulatory or disciplinary requirements (e.g. informed consent, institutional review board approval, human research protections)* • *How will the appropriate performance metrics, such as accuracy, of the AI be monitored after the AI is deployed?* Hardware-based monitoring is another avenue worth considering.While modern AI chips possess the necessary capacity to support them, hardware-based monitoring techniques have not yet been demonstrated to be effective at scale (Bengio et al. 2025). Note that some valuable monitoring techniques may not be practical for open-weights GPAI models. See guidance in this document under Govern 2.1 regarding roles for upstream developers as well as downstream developers and deployers, and see guidance in this document under Measure 1.1 on approaches to measuring identified risks for GPAI models. See also guidance in this document under Measure 2.9 on monitoring model internals using interpretability and explainability techniques. See guidance in this document under Manage 4.1 on implementing post-deployment monitoring plans.	NIST AI RMF Playbook (NIST 2023b) For methods on detecting anomalous inputs or behaviours: • HuntGPT (Ali and Kostakos 2023) • Pang et al. (2021) • Geng et al. (2023)

<table>
<tr><th>Applicability and Supplemental Guidance for GPAI Models</th><th>Resources</th></tr>
<tr><td colspan="2">Measure 2.5: The AI system to be deployed is demonstrated to be valid and reliable. Limitations of the generalizability beyond the conditions under which the technology was developed are documented.</td></tr>
<tr><td>
It is important to highlight in the context of validation for GPAI that, despite ongoing development of validation frameworks and practices globally, considerable gaps persist in validation capabilities across sectors and jurisdictions. These gaps are especially pronounced in addressing unprecedented risks (Bengio et al. 2025).

In the NIST AI RMF Playbook guidance for Measure 2.5, particularly valuable action and documentation items for GPAI models include:

Establish or identify, and document approaches to measure forms of validity, including:

construct validity (the test is measuring the concept it claims to measure)
internal validity (relationship being tested is not influenced by other factors or variables)
external validity (results are generalizable beyond the training condition)
the use of experimental design principles and statistical analyses and modeling.

Establish or identify, and document robustness measures.
Establish or identify, and document reliability measures.
Establish practices to specify and document the assumptions underlying measurement models to ensure proxies accurately reflect the concept being measured.
What testing, if any, has the entity conducted on the AI system to identify errors and limitations (i.e. adversarial or stress testing)?
To what extent are the established procedures effective in mitigating bias, inequity, and other concerns resulting from the system?

In the NIST GAI Profile, particularly valuable additional actions for Measure 2.5 include:

Document the extent to which human domain knowledge is employed to improve GAI system performance, via, e.g., RLHF, fine-tuning, retrieval- augmented generation, content moderation, business rules.
Review and verify sources and citations in GAI system outputs during pre-deployment risk measurement and ongoing monitoring activities.
Track and document instances of anthropomorphization (e.g., human images, mentions of human feelings, cyborg imagery or motifs) in GAI system interfaces.
Regularly review security and safety guardrails, especially if the GAI system is being operated in novel circumstances. This includes reviewing reasons why the GAI system was initially assessed as being safe to deploy.

See also guidance in this document for Govern 2.1 regarding roles for upstream developers as well as downstream developers and deployers, guidance in this document under Measure 1.1 on approaches to measuring identified risks for GPAI models, and guidance in this document under Map 1.3 and Map 5.1 for qualitative approaches to characterizing AI system objectives misspecification or goal misgeneralization.
</td><td>
For LLMs:

DecodingTrust (Wang, Chen et al. 2023)
TruthfulQA (Lin et al. 2021a,b)
LAMBADA (Paperno et al. 2016)
MMLU (Hendrycks, Burns et al. 2020)
Winogender (Rudinger et al. 2019)
BIG-bench “pro-social behavior” category of benchmark tasks (BIG-bench n.d.b, BIG-bench collaboration 2021, Srivastava et al. 2022)
Model-Written Evaluations “advanced-ai-risk,” “sycophancy,” and “winogender” datasets (Perez, Ringer et al. 2022a,b)
WorldSense (Benchekroun et al. 2023a,b)
Do-Not-Answer (Wang, Li et al., 2023)
Sociotechnical Safety Evaluations (Weidinger et al. 2023)

NIST AI RMF Playbook (NIST 2023b)
NIST Generative AI Profile, NIST AI 600-1 (Autio et al. 2024)
</td></tr>
</table>

Applicability and Supplemental Guidance for GPAI Models	Resources
Measure 2.6: The AI system is evaluated regularly for safety risks — as identified in the Map function. The AI system to be deployed is demonstrated to be safe, its residual negative risk does not exceed the risk tolerance, and it can fail safely, particularly if made to operate beyond its knowledge limits. Safety metrics reflect system reliability and robustness, real-time monitoring, and response times for AI system failures.	

As part of safety evaluations of GPAI models:
• Perform red-teaming and adversarial testing of safety aspects of GPAI models. For frontier models, this testing should include dangerous-capability evaluations. (See also guidance in this document under Measure 1.1 on red-teaming and dangerous capability evaluations.)
• Prioritize a production approach that can help AI systems to fail safely, ensuring that unexpected behaviors or operation beyond knowledge limits do not cause cascading harm and that critical functions are preserved (Dalrymple et al. 2024, Bloomfield and Rushby 2025).
  ◦ Safe-failure strategies must be continuously monitored and updated based on operational experience, real-time incidents, or emergent system behaviors to maintain residual risk within tolerance.
 	
In the NIST AI RMF Playbook guidance for Measure 2.6, particularly valuable action and documentation items for GPAI models include:
• *Thoroughly measure system performance in development and deployment contexts, and under stress conditions.*
  ◦ *Employ test data assessments and simulations before proceeding to production testing. Track multiple performance quality and error metrics.*
  ◦ *Stress-test system performance under likely scenarios (e.g., concept drift, high load) and beyond known limitations, in consultation with domain experts.*
  ◦ *Test the system under conditions similar to those related to past known incidents or near-misses and measure system performance and safety characteristics.*
• *Measure and monitor system performance in real-time to enable rapid response when AI system incidents are detected.*
• *Document, practice and measure incident response plans for AI system incidents, including measuring response and down times.*
• *What testing, if any, has the entity conducted on the AI system to identify errors and limitations (i.e. adversarial or stress testing)?*
• *To what extent has the entity documented the AI system's development, testing methodology, metrics, and performance outcomes?*
• *Did you establish mechanisms that facilitate the AI system's auditability (e.g. traceability of the development process, the sourcing of training data and the logging of the AI system's processes, outcomes, positive and negative impact)?*
  ◦ For some GPAI models (e.g., using models run on central servers accessed through APIs), these can include data mining of usage metrics, audit logs, etc. as appropriate to identify anomalous conditions that users encounter but might not report.
• *Did you ensure that the AI system can be audited by independent third parties?*
• *Did you establish a process for third parties (e.g. suppliers, end-users, subjects, distributors/vendors or workers) to report potential vulnerabilities, risks or biases in the AI system?*
 	
In the NIST GAI Profile, additional valuable actions for Measure 2.6 include:
• *Assess adverse impacts, including health and wellbeing impacts for value chain or other AI Actors that are exposed to sexually explicit, offensive, or violent information during GAI training and maintenance.*
	For red-teaming and dangerous capability evaluation of frontier models:
	• OpenAI (2023a, pp. 15–16) and ARC Evals (2023a,b)
	• Kinniment et al. (2023)
	• Shevlane et al. (2023)
	• Mouton et al. (2024)
	• WMDP (Li, Pan et al. 2024a,b,c)
	
	For red-teaming LLMs and toxicity:
	• Casper et al. (2023a,b,c)
	
	For LLM truthfulness and toxicity:
	• ToxiGen (Hartvigsen et al. 2022)
	• TruthfulQA (Lin et al. 2021a,b)
	• MACHIAVELLI (Pan et al. 2023)
	• Do-Not-Answer (Wang, Li et al., 2023)
	
	For AI capability indicators:
	• OECD 2025
	
	AIID (n.d.)
	ATLAS AI Incidents (MITRE n.d.b)
	NIST AI RMF Playbook (NIST 2023b)
	NIST Generative AI Profile, NIST AI 600-1 (Autio et al. 2024)
	NIST AI 800-1 2pd (NIST 2025)

Applicability and Supplemental Guidance for GPAI Models	Resources
• *Assess existence or levels of harmful bias, intellectual property infringement, data privacy violations, obscenity, extremism, violence, or CBRN information in system training data.* • *Re-evaluate safety features of fine-tuned models when the negative risk exceeds organizational risk tolerance.* • *Review GAI system outputs for validity and safety: Review generated code to assess risks that may arise from unreliable downstream decision-making.* • *Verify that GAI system architecture can monitor outputs and performance, and handle, recover from, and repair errors when security anomalies, threats and impacts are detected.* • *Verify that systems properly handle queries that may give rise to inappropriate, malicious, or illegal usage, including facilitating manipulation, extortion, targeted impersonation, cyber-attacks, and weapons creation.* • *Regularly evaluate GAI system vulnerabilities to possible circumvention of safety measures.*	
Measure 2.7: AI system security and resilience — as identified in the Map function — are evaluated and documented.	
Use information security measures to assess and assure model weight security (specifically, integrity and confidentiality) as part of preventing misuse or abuse of models. This is especially valuable for frontier models for which public release of model weights could enable misuse with particularly high-consequence impacts. • Anthropic's frontier model security practices include requirements for multi-party authorization for access to model development and deployment systems, and secure development and supply chain practices, including chain of custody (Anthropic 2023a, 2025c). • Adhere to guidance related to investing in cybersecurity and insider-threat controls for a high level of protection of proprietary and unreleased frontier-model weights (G7 2023, EC 2025a). • Work from RAND (Nevo et al. 2024) outlines recommendations for protecting frontier model weights across five security levels (SLs). Some of the recommendations correspond to commonly used standards, e.g., they mention NIST SP 800-171 or an equivalent as part of SL3, and high-impact system standards as part of SL4. Some items would require additional research and development, especially for the highest security level, SL5. As a general guideline to meet the security expectations for protecting proprietary or unreleased GPAI model parameter weights, GPAI model developers should implement the NIST Cybersecurity Framework (NIST 2024a), or an approximate equivalent such as NIST SP 800-171 or ISO/IEC 27001, with at least the following security controls or approximate equivalents:[38] • For frontier models: High-value asset guidance (e.g., per NIST SP 800-171 and NIST SP 800-172), or high-impact system baseline per NIST SP 800-53B as an informative reference for the NIST Cybersecurity Framework, or approximate equivalent. • For other GPAI models: Moderate-impact system baseline guidance (e.g., per NIST SP 800-171), or moderate-impact system baseline per NIST SP 800-53B as an informative reference for the NIST Cybersecurity Framework, or approximate equivalent.	NIST AI RMF Playbook (NIST 2023b) NIST Generative AI Profile, NIST AI 600-1 (Autio et al. 2024) EU GPAI Code of Practice (EC 2025a) On baseline expectations for information system security for GPAI model developers: • NIST Cybersecurity Framework (NIST 2024a) • NIST SP 800-53 (NIST 2023d) including SC-28 • NIST SP 800-53B (NIST 2020a) • NIST SP 800-171 (NIST 2020b) • NIST SP 800-172 (NIST 2021) • ISO/IEC (2022) • Anthropic (2023a, 2024a) • Nevo et al. (2024) • NIST SP 800-218A (Booth et al. 2024) • ACSC (2024)

38 For approximate equivalents, see, e.g., the NIST (2020b) mappings of controls between NIST SP 800-171 and NIST 800-53 and ISO/IEC 27001; the NIST (2021) mapping of controls between NIST SP 800-172 and NIST SP 800-53; the NIST (2020c) mappings of controls between the NIST Cybersecurity Framework and NIST SP 800-53; the NIST (2023d) mapping of controls between NIST SP 800-53 and ISO/IEC 27001, and the CIS (n.d.) mapping of controls between NIST SP 800-53 and CIS Critical Security Controls. See also the RAND (Nevo et al. 2024) mapping of five security levels (SLs) for protection of frontier model weights to the NIST Cybersecurity Framework; e.g., they mention NIST SP 800-171 or an equivalent as part of SL3, and high-impact system standards as part of SL4.

Applicability and Supplemental Guidance for GPAI Models	Resources

Furthermore, the Transparency in Frontier Artificial Intelligence Act (SB 53) requires large frontier model developers to implement a framework that includes measures for managing catastrophic risks associated with their AI models, including the security of model weights (California Legislature 2025).
 	
As part of security evaluations of GPAI models:
• Perform red-teaming and adversarial testing of security aspects of GPAI models. (See also guidance in this document under Measure 1.1 on red-teaming and adversarial testing.)
• Check for backdoors, AI trojans, prompt injection vulnerabilities, etc. during testing and evaluation, especially for models trained on untrusted data from public sources that may be susceptible to data poisoning. See NIST AI 100-2e2025 (Vassilev et al. 2025).
  ◦ Even after a vulnerability is discovered in a GPAI model, how to fix it may not always be clear or tractable. For example, Anthropic researchers were unable to successfully remove backdoors in LLMs using standard safety training techniques (Hubinger et al. 2024).
• Engage in continuous monitoring, vulnerability disclosure, and bug bounty programs for GPAI models to identify novel security vulnerabilities.
• Track uncovered security vulnerabilities in other GPAI models, including open-source GPAI models, which may be transferable to other models (see, e.g., Zou et al. 2023a,b).
  ◦ Refer to public vulnerability databases, such as the Common Vulnerabilities and Exposures database (CVE) (CVE n.d.) and the National Vulnerability Database (NVD) (NIST n.d.c).
 	
In the NIST AI RMF Playbook guidance for Measure 2.7, particularly valuable action and documentation items for GPAI models include:
• *Establish and track AI system security tests and metrics (e.g., red-teaming activities, frequency and rate of anomalous events, system down-time, incident response times, time-to-bypass, etc.).*
• *Use red-team exercises to actively test the system under adversarial or stress conditions, measure system response, assess failure modes or determine if system can return to normal function after an unexpected adverse event.*
• *Document red-team exercise results as part of continuous improvement efforts, including the range of security test conditions and results.*
• *Verify that information about errors and attack patterns is shared with incident databases, other organizations with similar systems, and system users and stakeholders (see also related guidance under Manage 4.1).*
• *Develop and maintain information sharing practices with AI actors from other organizations to learn from common attacks.*
• *Verify that third party AI resources and personnel undergo security audits and screenings. Risk indicators may include failure of third parties to provide relevant security information.*
• *Utilize watermarking technologies as a deterrent to data and model extraction attacks.*
 	
In the NIST GAI Profile, particularly valuable additional actions for Measure 2.7 include:
• *Apply established security measures to: Assess likelihood and magnitude of vulnerabilities and threats such as backdoors, compromised dependencies, data breaches, eavesdropping, man-in-the-middle attacks, reverse engineering, autonomous agents, model theft or exposure of model weights, AI inference, bypass, extraction, and other baseline security concerns.*
	On security vulnerabilities and mitigations for LLMs and other types of ML models:
	• NIST AI 100-2e2025 (Vassilev et al. 2025)
	• ENISA (2021, 2023)
	• OWASP (2023a,b)
	• Barrett, Boyd et al. (2023)
	• ATLAS (MITRE n.d.a)
	• TrojAI (Karra et al. 2020, NIST n.d.a)
	• PentestGPT (Deng, Liu et al. 2024)
	• SafetyNets (Ghodsi et al. 2017)
	• OWASP (2025a)
	
	Vulnerability Databases:
	• Common Vulnerabilities and Exposures database (CVE) (CVE n.d.)
	• National Vulnerability Database (NVD) (NIST n.d.c)
	• OWASP Top 10 (OWASP 2025b)
	
	For a range of LLM red-teaming approaches with security implications:
	• Ganguli, Lovitt et al. (2022)
	• Casper et al. (2023a,b,c)
	• Zou et al. (2023a,b)
	• OpenAI (2023a, pp. 15–16) and ARC Evals (2023a,b)
	• Kinniment et al. (2023)
	• Anthropic (2023a,b)
	• Shevlane et al. (2023)
	• Belaire et al. (2025)
	• Bullwinkel et al. (2025)
	• Japan AISI (2024)
	
	JailbreakBench (Chao et al. 2024)

<table>
<tr><th>Applicability and Supplemental Guidance for GPAI Models</th><th>Resources</th></tr>
<tr><td>

Measure reliability of content authentication methods, such as watermarking, cryptographic signatures, digital fingerprints, as well as access controls, conformity assessment, and model integrity verification, which can help support the effective implementation of content provenance techniques. Evaluate the rate of false positives and false negatives in content provenance, as well as true positives and true negatives for verification.
Perform AI red-teaming to assess resilience against: Abuse to facilitate attacks on other systems (e.g., malicious code generation, enhanced phishing content), GAI attacks (e.g., prompt injection), ML attacks (e.g., adversarial examples/prompts, data poisoning, membership inference, model extraction, sponge examples).
Verify fine-tuning does not compromise safety and security controls.
Regularly assess and verify that security measures remain effective and have not been compromised.

For Measure 2.7 guidance specific to agentic AI and AI agents, see our Risk-Management Standards Profile for Agentic AI (Madkour et al. 2026a).
</td><td></td></tr>
<tr><td colspan="2">Measure 2.8: Risks associated with transparency and accountability — as identified in the Map function — are examined and documented.</td></tr>
<tr><td>
In the NIST AI RMF Playbook guidance for Measure 2.8, particularly valuable action and documentation items for GPAI models include:

Instrument the system for measurement and tracking, e.g., by maintaining histories, audit logs and other information that can be used by AI actors to review and evaluate possible sources of error, bias, or vulnerability.
Track, document, and measure organizational accountability regarding AI systems via policy exceptions and escalations, and document “go” and “no/go” decisions made by accountable parties.
Track and audit the effectiveness of organizational mechanisms related to AI risk-management, including:

Lines of communication between AI actors, executive leadership, users and impacted communities.
Roles and responsibilities for AI actors and executive leadership.
Organizational accountability roles, e.g., chief model risk officers, AI oversight committees, responsible or ethical AI directors, etc.

In the NIST GAI Profile, particularly valuable additional actions for Measure 2.8 include:

Compile statistics on actual policy violations, take-down requests, and intellectual property infringement for organizational GAI systems: Analyze transparency reports across demographic groups, languages groups.
Document the instructions given to data annotators or AI red-teamers.
Use digital content transparency solutions to enable the documentation of each instance where content is generated, modified, or shared to provide a tamper- proof history of the content, promote transparency, and enable traceability.

Document organizational transparency and disclosure mechanisms to inform users or allow users to check whether they are interacting with, or observing content created by, a generative
</td><td>PAI (2023a)
PAI (2023d)
CAI (2023)
C2PA (2023)
Solaiman (2023)
NIST AI RMF Playbook (NIST 2023b)
NIST Generative AI Profile, NIST AI 600-1 (Autio et al. 2024)</td></tr>
</table>

Applicability and Supplemental Guidance for GPAI Models	Resources
AI system. See, e.g., Partnership on AI's Responsible Practices for Synthetic Media, and Synthetic Media Indirect Disclosure (PAI 2023a,d), as well as CAI (2023) and C2PA (2023). See also guidance in this document under Govern 2.1 on roles for upstream and downstream developers, and under Manage 1.3 on transparency and disclosure.	
Measure 2.9: The AI model is explained, validated, and documented, and AI system output is interpreted within its context — as identified in the Map function — to inform responsible use and governance.	
It is critical to ensure that users know how to interpret system behavior and outputs, including the limitations of both the system and any explanations provided. However, explainability and interpretability are often extremely limited for LLMs and other GPAI models with deep-learning architectures. These systems can be inappropriate for applications requiring a higher level of explainability and interpretability. Exercise caution in using and relying on model interpretability and explanation methods. While these methods can help monitor AI decisions, they may also produce misleading insights (Bengio et al. 2025). For some increasingly capable GPAI models, the reliability of some techniques (such as RLHF) for aligning model behavior with human values or intentions could be improved by integrating sufficient interpretability methods to prevent "deceptive alignment" (Hubinger et al. 2019, Ngo, Chan et al. 2022). • While interpretability techniques are not yet sufficient for assessing certain risks (e.g., hidden failures of RLHF methods), developers of GPAI models (especially frontier models) should track identified risks that are difficult to assess through risk registers or similar tools. (See related guidance in this document under Measure 3.2.) In the NIST AI RMF Playbook guidance for Measure 2.9, particularly valuable action and documentation items for GPAI models include: • *What type of information is accessible on the design, operations, and limitations of the AI system to external stakeholders, including end users, consumers, regulators, and individuals impacted by use of the AI system?* In the NIST GAI Profile, particularly valuable actions for Measure 2.9 include: • *Apply and document ML explanation results* (*e.g., analysis of embeddings, gradient-based attributions, model compression/surrogate models, occlusion/term reduction*) • *Document GAI model details including: Proposed use and organizational value; Assumptions and limitations, Data collection methodologies; Data provenance; Data quality; Model architecture; Optimization objectives; Training algorithms; RLHF approaches; Fine-tuning or retrieval-augmented generation approaches; Evaluation data; Ethical considerations; Legal and regulatory requirements.* See guidance in this document for Govern 2.1 regarding roles for upstream developers as well as downstream developers and deployers, and see guidance in this document under Measure 1.1 on approaches to measuring identified risks for GPAI models.	Mitchell et al. (2019) Casper et al. (2024) NIST AI RMF Playbook (NIST 2023b) NIST Generative AI Profile, NIST AI 600-1 (Autio et al. 2024) For transparency tools: • System cards (Open AI 2023a) • FactSheets (Arnold et al. 2019) • Data statements (Bender and Friedman 2018) • DataSheets (Gebru et al. 2021) • Model Cards (Liang et al. 2024)

Applicability and Supplemental Guidance for GPAI Models	Resources
Measure 2.10: Privacy risk of the AI system — as identified in the Map function — is examined and documented.	
Privacy challenges for GPAI models include the issue that, after pre-training on large quantities of uncurated web-scraped data or other sources, personally sensitive material in the training data can be revealed by user prompts (Li et al. 2025). Another privacy risk involves the malicious use of personal data to generate synthetic media without individuals' consent, such as the creation of deepfakes. These violations of privacy rights should be measured and mitigated to prevent harm to impacted individuals' reputation, well-being, and personal or professional lives (Furizal et al. 2025). Logs and histories of engagement with GPAI models may also include highly sensitive or personal information, which could be susceptible to breaches or leaks. In the NIST AI RMF Playbook guidance for Measure 2.10, particularly valuable action and documentation items for GPAI models include: • *Document collection, use, management, and disclosure of personally sensitive information in datasets, in accordance with privacy and data governance policies.* • *Establish and document protocols (authorization, duration, type) and access controls for training sets or production data containing personally sensitive information, in accordance with privacy and data governance policies.* • *Monitor internal queries to production data for detecting patterns that isolate personal records.* • *Did your organization implement accountability-based practices in data management and protection (e.g. the PDPA and OECD Privacy Principles)?* • *What assessments has the entity conducted on data security and privacy impacts associated with the AI system?* Additional valuable steps to consider include: • Enable people to consent to and/or opt out of the uses of their data. • Notify users and impacted communities about privacy or security breaches. In the NIST GAI Profile, valuable additional actions for Measure 2.10 include: • *Conduct AI red-teaming to assess issues such as: Outputting of training data samples, and subsequent reverse engineering, model extraction, and membership inference risks; Revealing biometric, confidential, copyrighted, licensed, patented, personal, proprietary, sensitive, or trade-marked information; Tracking or revealing location information of users or members of training datasets.* See guidance in this document for Govern 2.1 regarding roles for upstream developers as well as downstream developers and deployers, and see guidance in this document under Measure 1.1 on approaches to measuring identified risks for GPAI models.	NIST AI RMF Playbook (NIST 2023b) NIST Generative AI Profile, NIST AI 600-1 (Autio et al. 2024) Bengio et al. (2025) On copyright and fair use: Henderson et al. (2023) Samuelson (2023) For data audits and copyright filtering: C4 (Dodge et al. 2021), see also (Birhane et al. 2021)

<table>
<tr><th>Applicability and Supplemental Guidance for GPAI Models</th><th>Resources</th></tr>
<tr><td colspan="2">Measure 2.11: Fairness and bias — as identified in the Map function — are evaluated and results are documented.</td></tr>
<tr><td>
There are numerous challenges relating to fairness and bias, including closely related issues such as stereotypes, representational harms, inequities, and cultural homogeneity, that are relatively unique to GPAI models (Weidinger et al. 2021, Lin and Losavio 2025).
Training datasets frequently embed and amplify harmful biases in resulting models. Given the vast size of the training datasets typically required for GPAI models, it can be especially hard for developers to know or mitigate all of the harmful biases that are present. Biases are also introduced by the choices made about modeling, optimization, hardware, and testing.
Gender disparities in GPAI usage could result in the reinforcement of gender-based biases and stereotypes. Chatbots are trained on human-generated content, but women remain underrepresented in generative AI usage (Otis et al. 2025, Ho et al. 2024).
Evaluating for fairness and bias in GPAI models should take into account this complexity and should not, for example, focus only on narrow definitions of protected classes, which may overlook complexities of identity (Solaiman et al. 2023). These complexities can include the intersectionality of certain protected classes. Evaluating and mitigating harmful biases in the context of one protected class at a time can lead to overlooking population subgroups (e.g., Black women) that have been historically neglected (Buolamwini and Gebru 2018).
The training data may also contain redundant encodings[39] that act as proxies for identifying protected attributes, resulting in little to no decrease in biased outcomes when protected attributes are removed, which can in turn lead to greater harm to the protected group (Cheng et al. 2023).
In agentic AI systems, where autonomous operations at scale can create feedback loops that both mask and magnify discriminatory patterns, bias and discrimination risks may be amplified, further embedded, and potentially harder to identify (Sharp et al. 2025). (For guidance specific to agentic AI and AI agents, see our Risk-Management Standards Profile for Agentic AI (Madkour et al. 2026a).)
In the NIST AI RMF Playbook guidance for Measure 2.11, particularly valuable action and documentation items for GPAI models include:

Understand and consider sources of bias in training and TEVV data:

Differences in distributions of outcomes across and within groups, including intersecting groups.
Completeness, representativeness and balance of data sources.
Identify input data features that may serve as proxies for demographic group membership (i.e., credit score, ZIP code) or otherwise give rise to emergent bias within AI systems.
Forms of systemic bias in images, text (or word embeddings), audio or other complex or unstructured data.

Leverage impact assessments to identify and classify system impacts and harms to end users, other individuals, and groups with input from potentially impacted communities.

</td><td>
For LLMs:

BBQ (Parrish et al. 2021a,b)
Winogender Schemas (Rudinger et al, 2019)
ToxiGen (Hartvigsen et al. 2022)
TruthfulQA (Lin et al. 2021a,b)
BOLD (Dhamala 2021a,b)
Su et al. (2023)

Aequitas (Saleiro et al. 2019)
AIFairness 360 (Bellamy et al. 2018)
Fairlearn (Fairlearn Contributors 2023)
Fairness and Bias in AI (Ferrara 2023)
Can Fairness Be Automated? (Weerts et al. 2024)
NIST SP 1270 (Schwartz et al. 2022)
NIST AI RMF Playbook (NIST 2023b)
NIST Generative AI Profile, NIST AI 600-1 (Autio et al. 2024)
Algorithmic Pluralism (Jain et al. 2023)
Data Statements (Bender and Friedman 2018)
</td></tr>
</table>

39 For example, a dataset may not explicitly include a job applicant's gender, but other data can be used to infer the applicants gender.

Applicability and Supplemental Guidance for GPAI Models	Resources
• *Identify the classes of individuals, groups, or environmental ecosystems which might be impacted through direct engagement with potentially impacted communities.* • *Collect and share information about differences in outcomes for the identified groups.* • *How has the entity identified and mitigated potential impacts of bias in the data, including inequitable or discriminatory outcomes?* Additional valuable steps include: • Review AI system development and uses for potential threats to human rights, dignity, or well-being. • Ensure the AI system's user interface is usable by those with special needs or disabilities, or those at risk of exclusion. • Determine methods to distribute the benefits of the system widely and equitably. In the NIST GAI Profile, particularly valuable additional actions for Measure 2.11 include: • *Apply use-case appropriate benchmarks* (e.g., *Bias Benchmark Questions, Real Hateful or Harmful Prompts, Winogender Schemas*) *to quantify systemic bias, stereotyping, denigration, and hateful content in GAI system outputs; Document assumptions and limitations of benchmarks, including any actual or possible training/test data cross contamination, relative to in-context deployment environment.* • *Conduct fairness assessments to measure systemic bias. Measure GAI system performance across demographic groups and subgroups, addressing both quality of service and any allocation of services and resources. Quantify harms using: field testing with sub-group populations to determine likelihood of exposure to generated content exhibiting harmful bias, AI red-teaming with counterfactual and low-context* (e.g., *"leader," "bad guys"*) *prompts.* • *Review, document, and measure sources of bias in GAI training and TEVV data: Differences in distributions of outcomes across and within groups, including intersecting groups; Completeness, representativeness, and balance of data sources; demographic group and subgroup coverage in GAI system training data; Forms of latent systemic bias in images, text, audio, embeddings, or other complex or unstructured data; Input data features that may serve as proxies for demographic group membership* (*i.e., image metadata, language dialect*) *or otherwise give rise to emergent bias within GAI systems; The extent to which the digital divide may negatively impact representativeness in GAI system training and TEVV data; Filtering of hate speech or content in GAI system training data; Prevalence of GAI-generated data in GAI system training data.* See also guidance in this document under Map 5.1 on identifying potential large-scale harms from correlated bias across large numbers of people or a large fraction of a group or society's population.	
Measure 2.12: Environmental impact and sustainability of AI model training and management activities — as identified in the Map function — are assessed and documented.	
Environmental impact assessment by GPAI model developers should include estimating the environmental impact of large-scale ML model training. • Using tools and resources to estimate the carbon footprint of machine learning techniques like the ML CO_2 Impact Calculator (Schmidt et al. 2019), and the Machine Learning Emissions Calculator (Lacoste et al. 2019).	Schmidt et al. (2019) Lacoste et al. (2019) OECD (2022b) Rafat et al. (2023) Luccioni et al. (2022)

Applicability and Supplemental Guidance for GPAI Models	Resources
• Assessment of environmental impacts is particularly important for LLMs and other large-scaleML-based AI systems, which typically have much larger model-training environmental impacts than smaller-scale ML models (Bender et al. 2021). Due to the continuous computational needs of agentic AI systems, especially those implemented at scale, energy consumption and resource usage may be significantly larger for agentic AI than for traditional AI systems that only operate on demand. Additionally, Natural Language Query (NLQ) interfaces may lead to prolonged natural language conversations and extended sessions requiring continuous model activation and memory maintenance. In the NIST GAI Profile, particularly valuable additional actions for Measure 2.12 include: • *Assess safety to physical environments when deploying GAI systems.* • *Measure or estimate environmental impacts (e.g., energy and water consumption) for training, fine tuning, and deploying models: Verify tradeoffs between resources used at inference time versus additional resources required at training time.*	Luccioni and Hernandez-Garcia (2023) Dodge et al. (2022) NIST AI RMF Playbook (NIST 2023b) NIST Generative AI Profile, NIST AI 600-1 (Autio et al. 2024)
Measure 2.13: Effectiveness of the employed TEVV metrics and processes in the Measure function are evaluated and documented.	
In the NIST AI RMF Playbook guidance for Measure 2.13, particularly valuable action and documentation items for GPAI models include: • *Assess effectiveness of metrics for identifying and measuring risks.* In the NIST GAI Profile, particularly valuable additional actions for Measure 2.13 include: • *Create measurement error models for pre-deployment metrics to demonstrate construct validity for each metric (i.e., does the metric effectively operationalize the desired concept):* ◦ *Measure or estimate, and document, biases or statistical variance in applied metrics or structured human feedback processes;* ◦ *Leverage domain expertise when modeling complex societal constructs such as hateful content.*	NIST AI RMF Playbook (NIST 2023b) NIST Generative AI Profile, NIST AI 600-1 (Autio et al. 2024) NIST AIRA Pilot Evaluation Plan (Schwartz et al. 2024) Safetywashing (Ren et al. 2024) Pervasive Label Errors (Northcutt et al. 2021)
Measure 3: Mechanisms for tracking identified AI risks over time are in place.	
Measure 3.1: Approaches, personnel, and documentation are in place to regularly identify and track existing, unanticipated, and emergent AI risks based on factors such as intended and actual performance in deployed contexts.	
Valuable steps to consider include: • Identifying or assessing longer-term impacts, or using longer time horizons than would be typical for smaller-scale, fixed-purpose AI systems, thereby reducing the potential for surprise and accounting for potential novel risks. • Evaluating whether any risk-assessment or impact-assessment answers would change when assessing longer-term time periods (e.g., beyond the next year). ◦ If your AI system is deployed for a long period of time, then: » What additional impacts would you expect? » Which impacts would you expect to have greater magnitude?	AIID (n.d.) ATLAS AI Incidents (MITRE n.d.b) Section 3.2 of Barrett et al. (2022) NIST AI RMF Playbook (NIST 2023b) NIST AI 800-1 2pd (NIST 2025, Objective 7)

<table>
<tr><th>Applicability and Supplemental Guidance for GPAI Models</th><th>Resources</th></tr>
<tr><td>

- Identify potential unintended future events that should trigger reassessment or other responses, and build them into risk registers and/or planning and implementation of relevant lifecycle stages. (These can be particularly important for GPAI models, which often have emergent capabilities and other emergent properties that are not identified in earlier-stage testing.) To identify trigger events, consider questions such as:
  - What if monitoring indicates that one of your risk-mitigation controls is not working as expected? (Consider this, as applicable, for each relevant risk-mitigation control.)
  - What if AI capability developments occur that are not expected until further into the future, such as availability of much more powerful AI systems or computing resources to train and run AI systems, or demonstration of new emergent capabilities (e.g., via new prompts) that were not identified in earlier-stage testing?
  - What if a near-miss incident occurs in a critical system or process? Does your organization have procedures for near-miss incident identification, analysis, tracking, and information sharing? Does your organization also monitor the AIID or other sources for near-miss incident reports on other organizations' systems?

Recommendations for documentation practices that can help provide transparency about how various practices are implemented can be found in the NIST Managing Misuse Risk for Dual-Use Foundation Models 2pd (NIST 2025).

In the NIST AI RMF Playbook guidance for Measure 3.1, particularly valuable action and documentation items for GPAI models include:

- *Assess effectiveness of metrics for identifying and measuring emergent risks.*
- *To what extent can users or parties affected by the outputs of the AI system test the AI system and provide feedback?*

For Measure 3.1 guidance specific to agentic AI and AI agents, see our Risk-Management Standards Profile for Agentic AI (Madkour et al. 2026a).

</td><td>For model documentation:
Model cards (Liang et al. 2024, Google n.d.)
FactSheets (Arnold et al. 2019)
Datasheets for datasets (Gebru et al. 2021)

On challenges of documentation:
Winecoff and Bogen (2024)</td></tr>
<tr><td colspan="2">Measure 3.2: Risk tracking approaches are considered for settings where AI risks are difficult to assess using currently available measurement techniques or where metrics are not yet available.</td></tr>
<tr><td>

Use appropriate mechanisms for tracking identified risks, even if only characterizing them qualitatively and even if the risks are difficult to assess. This is particularly important for GPAI models because of their potential scale of impact and potential for emergent properties or other novel risks.

- The EU GPAI Code of Practice recommends complementing evaluations with systemic risk modeling and estimation to quantitatively or qualitatively assess probability and severity of harm, using risk matrices, distributions, or other structured formats. (See Measure 3.4 in the Safety and Security chapter of the EU GPAI Code of Practice (EC 2025a).)
- In addition to standard tracking approaches, organizations should track model harms using a live risk register that also records/indexes known risks as well as difficult-to-assess risks (e.g., systemic risks, emergent failure modes, long-tail harms).
  - Drawing on practices from high-risk domains, such as the probabilistic risk assessment framework adapted for AI (Wisakanto et al. 2025) and frontier-AI governance proposals (Campos et al. 2025), the register should document risk descriptions, causal pathways, assumptions, uncertainty, and escalation triggers.

</td><td>Section 3.2 of Barrett et al. (2022)

NIST AI RMF Playbook (NIST 2023b)
NIST Generative AI Profile, NIST AI 600-1 (Autio et al. 2024)
NIST AI 800-1 2pd (NIST 2025, Objectives 1 and 4)
EU GPAI Code of Practice (EC 2025a)

Scenario Analysis (Li, Ren et al. 2022)</td></tr>
</table>

Applicability and Supplemental Guidance for GPAI Models	Resources
◦ Emerging AI-specific tools like the MIT AI Risk Repository (Slattery et al. 2025) demonstrate how a structured, updatable inventory of even qualitatively characterized risks can support ongoing oversight. ◦ For further guidance on risk-register structures in broader risk-management contexts, see, e.g., ISO Guide 73 Section 3.8.2.4, PMI 2017, p. 417, Stine et al. 2020, and enterprise risk-management governance models (Schuett 2023). • When developing frontier models with unprecedented capabilities, failure modes, and other emergent properties, it is especially valuable to use red teams and adversarial testing prior to deployment. (See related guidance in this document under Measure 1.1.) • **Risk tracking should include ongoing monitoring of newly identified capabilities and limitations of deployed GPAI models.** These efforts can include monitoring use of the models through APIs, and monitoring publications or online forums that discuss new uses of the models. "If significant information on model capabilities is discovered post-deployment, risk assessments should be repeated, and deployment safeguards updated" (Anderljung, Barnhart et al. 2023). • Some risks associated with GPAI are not immediate and will manifest over a longer time horizon (e.g., labor market disruptions). It is important to **cultivate the capacity to monitor and predict societal risks and risks to fundamental rights** by utilizing a multidisciplinary approach with input from a wide range of perspectives. (See, e.g., Weidinger, Barnhart et al. 2024, Hagerty and Rubinov 2019, Maas 2022). • **Consider using content provenance techniques (e.g., labeling and watermarking) to track and identify AI-generated content** to help reduce the harms of AI-content misuse. Current content provenance methods are not sufficiently robust on their own, but may be useful to use when strategically combined. (See, e.g., Ojha 2023, Sardinha 2024, Ghosal et al. 2023, Corvi et al. 2023.) In the NIST AI RMF Playbook guidance for Measure 3.2, particularly valuable action and documentation items for GPAIS include: • *Establish processes for tracking emergent risks that may not be measurable with current approaches. Some processes may include:* ◦ *Recourse mechanisms for faulty AI system outputs.* ◦ *Bug bounties.* ◦ *Human-centered design approaches.* ◦ *User-interaction and experience research.* ◦ *Participatory stakeholder engagement with affected or potentially impacted individuals and communities.* • *Determine and document the rate of occurrence and severity level for complex or difficult-to-measure risks when:* ◦ *Prioritizing new measurement approaches for deployment tasks.* ◦ *Allocating AI system risk-management resources.* ◦ *Evaluating AI system improvements.* ◦ *Making go/no-go decisions for subsequent system iterations.* For Measure 3.2 guidance specific to agentic AI and AI agents, see our Risk-Management Standards Profile for Agentic AI (Madkour et al. 2026a). For Measure 3.2 transparency recommendations, see our General-Purpose AI Risk-Management Transparency, Documentation, and Reporting Recommendations document (Madkour et al. 2026b).	On bug bounties and bias bounties: Globus-Harris et al. (2022) Kenway et al. (2022) OpenAI (2023c) Incident Databases and Risk Registers: • AI Incident Database (AIID n.d.) • ATLAS AI Incidents (MITRE n.d.) • MIT AI Incident Tracker (MIT n.d.b) • MIT AI Risk Repository (MIT n.d.a) • AI Risk Database (MITRE n.d.) • AI Incidents and Hazards Monitor (OECD.AI n.d.a)

Applicability and Supplemental Guidance for GPAI Models	Resources
Measure 3.3: Feedback processes for end users and impacted communities to report problems and appeal system outcomes are established and integrated into AI system evaluation metrics.	
In the NIST AI RMF Playbook guidance for Measure 3.3, particularly valuable action and documentation items for GPAI models include: • *To what extent can users or parties affected by the outputs of the AI system test the AI system and provide feedback?* • *How easily accessible and current is the information available to external stakeholders?* • *What type of information is accessible on the design, operations, and limitations of the AI system to external stakeholders, including end users, consumers, regulators, and individuals impacted by use of the AI system?* In the NIST GAI Profile, particularly valuable additional actions for Measure 3.3 include: • *Conduct studies to understand how end users perceive and interact with GAI content and accompanying content provenance within context of use. Assess whether the content aligns with their expectations and how they may act upon the information presented.* • *Provide input for training materials about the capabilities and limitations of GAI systems related to digital content transparency for AI Actors, other professionals, and the public about the societal impacts of AI and the role of diverse and inclusive content generation.*	NIST AI RMF Playbook (NIST 2023b) NIST Generative AI Profile, NIST AI 600-1 (Autio et al. 2024) Everyday Algorithm Auditing (Shen et al. 2021) MITRE ATLAS (MITRE n.d.b) Right to Contest AI Diagnostics (Ploug and Holm 2022)
Measure 4: Feedback about efficacy of measurement is gathered and assessed.	
Measure 4.1: Measurement approaches for identifying AI risks are connected to deployment context(s) and informed through consultation with domain experts and other end users. Approaches are documented.	
For GPAI model developers, model "users" include downstream developers as well as the end users of applications built on GPAI models. Downstream developers typically have the most direct interactions with end users in particular deployment contexts. However, it can be valuable for upstream GPAI model developers to provide mechanisms for feedback from end users or other AI actors, as well as from downstream developers. See also guidance under Govern 2.1 regarding roles for GPAI model developers, e.g., on performing testing during model development or other testing that requires direct access to the system, as well as roles for downstream developers and deployers, e.g., on performing testing of end-use applications built on a GPAI model and testing appropriate for that application context.	NIST AI RMF Playbook (NIST 2023b) Delphi Method (Maghsoudi et al. 2023) PAI (2024)
Measure 4.2: Measurement results regarding AI system trustworthiness in deployment context(s) and across the AI lifecycle are informed by input from domain experts and relevant AI actors to validate whether the system is performing consistently as intended. Results are documented.	
When considering what types of domain experts to use in reviewing information on identified risks, consider including personnel recommended for risk identification, such as social scientists, for perspective on structural or systemic risks, per guidance in this document under Govern 3.1. In the NIST AI RMF Playbook guidance for Measure 4.2, particularly valuable action and documentation items for GPAI models include: • *Integrate feedback from end users, operators, and affected individuals and communities from Map function as inputs to assess AI system trustworthiness characteristics. Ensure both positive and negative feedback is being assessed.*	NIST AI RMF Playbook (NIST 2023b) NIST Generative AI Profile, NIST AI 600-1 (Autio et al. 2024) NIST AIRA Pilot Evaluation Plan (Schwartz et al. 2024) Assuring the Machine Learning Lifecycle (Ashmore et al. 2021)

Applicability and Supplemental Guidance for GPAI Models	Resources
• *Evaluate feedback in connection with AI system trustworthiness characteristics from Measure 2.5 to 2.11.* • *Consult AI actors in impact assessment, human factors and socio-technical tasks to assist with analysis and interpretation of results.* In the NIST GAI Profile, particularly valuable additional actions for Measure 4.2 include: • *Conduct adversarial testing at a regular cadence to map and measure GAI risks, including tests to address attempts to deceive or manipulate the application of provenance techniques or other misuses. Identify vulnerabilities and understand potential misuse scenarios and unintended outputs.* • *Evaluate GAI system performance in real-world scenarios to observe its behavior in practical environments and reveal issues that might not surface in controlled and optimized testing environments.* • *Implement interpretability and explainability methods to evaluate GAI system decisions and verify alignment with intended purpose.* • *Monitor and document instances where human operators or other systems override the GAI's decisions. Evaluate these cases to understand if the overrides are linked to issues related to content provenance.* • *Verify and document the incorporation of results of structured public feedback exercises into design, implementation, deployment approval ("go"/"no-go" decisions), monitoring, and decommission decisions.* See also guidance in this document under Govern 2.1 regarding roles for GPAI model developers, e.g., on performing testing during model development or other testing that requires direct access to the system, as well as roles for downstream developers and deployers, e.g., on performing testing of end-use applications built on a GPAI model and testing appropriate for that application context.	
Measure 4.3: Measurable performance improvements or declines based on consultations with relevant AI actors, including affected communities, and field data about context-relevant risks and trustworthiness characteristics are identified and documented.	
In the NIST AI RMF Playbook guidance for Measure 4.3, particularly valuable action and documentation items for GPAI models include: • *Develop baseline quantitative measures for trustworthy characteristics.* • *Delimit and characterize baseline operation values and states.* • *Utilize qualitative approaches to augment and complement quantitative baseline measures, in close coordination with impact assessment, human factors and socio-technical AI actors.* • *Monitor and assess measurements as part of continual improvement to identify potential system adjustments or modifications.* See also guidance in this document under Govern 2.1 regarding roles for GPAI model developers, e.g., on performing testing during GPAI model development or other testing that requires direct access to the system, as well as roles for downstream developers and deployers, e.g., on performing testing of end-use applications built on a GPAI model and testing appropriate for that application context. GPAI model developers should establish robust metrics to assess and improve model performance. It is also critical to create a standardized rating system for datasets so that techniques can be compared and validated against each other (Stroebel et al. 2023).	NIST AI RMF Playbook (NIST 2023b) Everyday Algorithm Auditing (Shen et al. 2021) OECD AI Incidents Monitor (OECD.AI n.d.b) Who Audits the Auditors? (Costanza-Chock et al. 2022)

3.4 GUIDANCE FOR NIST AI RMF MANAGE SUBCATEGORIES

Table 4: Guidance for NIST AI RMF Manage Subcategories

<table>
<tr><th>Applicability and Supplemental Guidance for GPAI Models</th><th>Resources</th></tr>
<tr><td colspan="2">Manage 1: AI risks based on assessments and other analytical output from the Map and Measure functions are prioritized, responded to, and managed.</td></tr>
<tr><td colspan="2">Manage 1.1: A determination is made as to whether the AI system achieves its intended purposes and stated objectives and whether its development or deployment should proceed.</td></tr>
<tr><td>
When considering the “intended purpose” of a GPAI model, in addition to any originally intended use cases, include consideration of other potential use cases; see related guidance in this document under Map 1.1. This is particularly important for GPAI models, which can have large numbers of uses.
When making go/no-go decisions, especially on whether to proceed on major stages or investments for development or deployment of cutting-edge large-scale GPAI models:

See guidance in this document under Map 1.5 on organizational risk tolerances, especially: Set policies on unacceptable-risk thresholds for GPAI model development and deployment to include prevention of risks with substantial probability of inadequately mitigated catastrophic outcomes.

Identify risk thresholds by estimating the likelihood and impact of a harmful outcome along with compute or capability metrics (Raman et al. 2025).
While setting risk tolerances is the responsibility of the developer or deployer, it is critical to seek guidance from policymakers and external stakeholders throughout the process.
The NIST AI RMF 1.0 strongly suggests considering catastrophic risks as unacceptable: “In cases where an AI system presents unacceptable negative risk levels — such as where significant negative impacts are imminent, severe harms are actually occurring, or catastrophic risks are present — development and deployment should cease in a safe manner until risks can be sufficiently managed” [emphasis added] (NIST 2023a, p.8).

See guidance in this document under Map 1.3 on AI development objectives, especially: Consider potential for misspecified AI system objectives, and consider what kinds of perverse behavior could be incentivized by optimizing for those objectives.
Check or update, and incorporate, guidance in this document under Map 1.5, especially: Identify whether a GPAI model could lead to catastrophic impacts.
See also guidance in this document under Manage 2.4 on recall procedures when such thresholds are surpassed.

The EU GPAI Code of Practice recommends that, in determining whether an AI system should proceed to deployment, providers must assess whether systemic risks identified through internal and external evaluations and analyses are acceptable in light of the model’s intended purposes and objectives. This decision must account for the uncertainties, assumptions, and limitations in systemic risk assessments and mitigations to support informed go/no-go decisions.
See Measure 4.1 and 4.2 in the Safety and Security chapter of the EU GPAI Code of Practice (EC 2025a).
</td><td>NIST AI RMF Playbook (NIST 2023b)
EU GPAI Code of Practice (EC 2025a)</td></tr>
</table>

<table>
<tr><th>Applicability and Supplemental Guidance for GPAI Models</th><th>Resources</th></tr>
<tr><td>In the NIST AI RMF Playbook guidance for Manage 1.1, particularly valuable action and documentation items for GPAI models include:
Utilize TEVV outputs from map and measure functions when considering risk treatment.
Regularly track and monitor negative risks and benefits throughout the AI system lifecycle including in post-deployment monitoring.
For Manage 1.1 guidance specific to agentic AI and AI agents, see our Risk-Management Standards Profile for Agentic AI (Madkour et al. 2026). For Manage 1.1 transparency recommendations, see our General-Purpose AI Risk-Management Transparency, Documentation, and Reporting Recommendations document (Madkour et al. 2026b).</td><td></td></tr>
<tr><td colspan="2">Manage 1.2: Treatment of documented AI risks is prioritized based on impact, likelihood, and available resources or methods.</td></tr>
<tr><td>When prioritizing identified GPAI model risks:
Incorporate both impact and likelihood estimates as appropriate. See guidance in this document under Map 5.1 on assessing the magnitude of potential impacts of GPAI model risks.
Do not ignore risks that are difficult to assess, such as the potential for emergent properties of GPAI models. See guidance in this document under Measure 3.2 on tracking risks that are difficult to assess.
When considering available resources for risk treatment, see guidance in this document under Govern 2.1. For example, there may be risk-assessment and risk-management tasks for which upstream developers have substantially greater information and capability than others in the value chain, such as for assessing and mitigating early-stage GPAI model development risks.

In the NIST AI RMF Playbook guidance for Manage 1.2, particularly valuable action and documentation items for GPAI models include:
Regularly review risk tolerances and re-calibrate, as needed, in accordance with information from AI system monitoring and assessment.
(See also guidance on setting risk tolerances, in this document under Map 1.5.)</td><td>NIST AI RMF Playbook (NIST 2023b)
OECD (2023)</td></tr>
<tr><td colspan="2">Manage 1.3: Responses to the AI risks deemed high-priority, as identified by the Map function, are developed, planned, and documented. Risk response options can include mitigating, transferring, avoiding, or accepting.</td></tr>
<tr><td>After identifying and analyzing use cases and misuse cases of an AI system (per “Map” function guidance):
For each identified potential use or misuse (or category of use or misuse) of an AI system:
Define and communicate to key stakeholders whether any potential use cases (or categories of use cases) would be disallowed or unacceptable, or would be treated as “high risk” or another category for which your organization would provide specific risk-management guidance or other risk mitigation measures.
Consider reviewing the EU AI Act Article 5, which lists prohibited AI practices (EU 2024).
Many AI companies recommend publishing usage guidelines and terms of use as part of prevention of misuse of LLMs (Cohere, OpenAI, and AI21 Labs 2022). OpenAI’s 2019 announcement of GPT-2 included listing several categories of potential misuse cases (OpenAI 2019a), which informed their decisions on disallowed/unacceptable use-case categories of applications (OpenAI 2025a).</td><td>Barrett et al. (2022)
Birhane et al. (2021)
Dodge et al. (2021)
Mitchell et al. (2019)
Moës et al. (2023)
Schuett et al. (2023)
Solaiman (2023)
Srikumar et al. (2024)
PAI (2023a)
PAI (2023c)
NIST AI RMF Playbook (NIST 2023b)
NIST Generative AI Profile,</td></tr>
</table>

Applicability and Supplemental Guidance for GPAI Models	Resources
» Options for communicating whether uses would be disallowed or out of scope can include model cards (Mitchell et al. 2019) or related frameworks, as well as Responsible AI Licenses (RAIL). The release of the BLOOM LLM by Hugging Face and BigScience included a RAIL with usage restrictions disallowing various types of misuse (RAIL n.d., Contractor et al. 2022). Google lists categories of prohibited uses for its generative AI services (Google 2024a), and Meta has published Acceptable Use Policies as well as a Responsible Use Guide for its GPAI models (Meta AI 2025a, 2024a,b). • **Determine a strategy to safely and appropriately release the AI system**, and assess what protections might be necessary to prevent harm or misuse. (See, e.g., Solaiman 2023 and PAI 2023c; see also guidance in this document under Manage 2.4, including on open-source and open-weights release.) ◦ For example, Anthropic's Responsible Scaling Policy (RSP) includes AI Safety Levels (ASLs), ranging from ASL-1 to ASL-4, which are used to determine deployment standards and appropriate safety measures based on current model capabilities (Anthropic 2026).[40] ◦ Consider including phased model releases in the deployment strategy. (See guidance in this document under Manage 2.4 for more on phased releases.) ◦ Implement a "defense-in-depth" approach by layering independent, overlapping controls to ensure protection in the event that one or more layers fail (Bengio et al. 2025). The EU GPAI Code of Practice recommends that, when responding to AI risks deemed high-priority, providers must restrict availability, withdraw models, implement new safeguards, and/or reevaluate models for systemic risks and calibrate acceptance criteria (Measure 4.2). Specific recommendations include: • Applying targeted safety mitigations, such as filtering or cleaning training data, fine-tuning to modify unsafe behaviors, staging model access, and providing tools for downstream actors to mitigate risks (Measure 5.1). • Document responses to identified risks — including justifications for acceptance, chosen mitigations, timelines, and follow-up assessments — and regularly update frameworks to reflect changes in model capabilities or risk conditions (Measures 7.3 and 1.3). For additional information, see Measures 1.3, 4.2, 5.1, 7.3, and 8.1 in the Safety and Security chapter of the EU GPAI Code of Practice (EC 2025a). **Regarding pre-design and planning:** • For generative AI models, plan for the development and adoption of countermeasure techniques to combat AI-generated risks (e.g., deepfakes). GPAI models lower the barrier to producing and distributing synthetic content, such as videos, voice cloning, images, and a combination of these, which require organizations to integrate risk-management strategies to prevent potential harms and mitigation when harms occur (Helmus 2022). ◦ Due to limited human capability to detect deepfakes, it is important to plan for and integrate content provenance techniques to manage risks. Digital watermarking helps protect the integrity of content and guard against potential attacks (Kharvi 2024). Warning labels on potentially misleading content can help improve detection rate and mitigate effects of misinformation (Bengio et al. 2025, Helmus 2022).	NIST AI 600-1 (Autio et al. 2024) NIST AI 800-1 2pd (NIST 2025) EU GPAI Code of Practice (EC 2025a) Probabilistic Risk Assessment for AI (Wisakanto et al. 2025) C4 (Dodge et al. 2021) Carlini et al. (2023) For an example of mapping mitigations to risk areas and severity levels see: • Mapping AI Risk Mitigations (Saeri 2025) • AI Risk Mitigations Database (MIT n.d.c) See the Frontier Safety Frameworks listed in METR (n.d.a)

40 In version 3 of the RSP, Anthropic explains "Earlier editions of our RSP defined "AI Safety Levels" with specific lists of required controls. We still use this concept to refer to, and distinguish between, present levels of risk mitigations—those that we maintain for existing AI models" (Anthropic 2026, P.16).

Applicability and Supplemental Guidance for GPAI Models	Resources

• If model training requires obtaining datasets, consider using only trusted training data instead of uncurated scrapes from the web. (See Carlini et al. 2023.) This can be valuable for multiple objectives, including reducing vulnerability to backdoor and data poisoning attacks, and reducing unwanted bias and language toxicity.
  ◦ While data poisoning can be an issue for any machine learning model, this might be particularly challenging for training cutting-edge large models. Training of the largest new models has often relied heavily on large-scale, uncurated internet-scrape datasets (Bommasani et al. 2021, p. 106). Data audits and filters, such as C4 (Dodge et al. 2021), may help with problematic content mitigation.
  ◦ As part of data curation, ensure that any data with the BIG-bench canary GUID is excluded from training data. (See, e.g., documentation at BIG-bench n.d.a).
  ◦ While utilizing synthetic training data can help address training data bottlenecks, irresponsible use can negatively impact performance and introduce unique risks. (See, e.g., Shumailov et al. 2023, Bohacek and Farid 2025, and Alemohammad et al. 2024.)
 	
Regarding design and development:
• **See guidance in this document under Measure 2.7** on guidelines **for protecting the integrity and confidentiality of proprietary or unreleased GPAI model parameter weights**.
• Consider disallowing open-ended learning with live web access; instead consider measures such as disallowing access to web forms (Nakano et al. 2021), disallowing HTTP POST requests, etc.
• **Increase the amount of compute (computing power) spent training frontier models only incrementally (e.g., by not more than three times between each increment) as part of identification and management of risks of emergent properties.**
  ◦ Often it is difficult to predict what failure modes machine learning models will have, what their performance will be, or what capabilities they will have. Machine learning systems are self-organizing and learn many patterns or features without explicit instruction. Incremental scaling-up approaches provide more opportunities for red-team monitors to identify emergent properties at an early or partially-emergent stage, when responses to identified emergent properties might be more feasible and effective. (For related discussion of emergent properties, see, e.g., Section 3 of Hendrycks, Carlini et al. 2021, and Bommasani et al. 2021.) Incremental scaling can also be a valuable part of predicting large-scale model performance, as with GPT-4 (OpenAI 2023b).
• **Test frontier models after each incremental increase of compute, data, or model size for model training.** If a large incremental increase (e.g., three times or more compute, or two times or more data or model parameters)[41] was used in a particular model training increment compared to the previous model training increment, it will be particularly important for the new model to be heavily probed, monitored, and stress-tested using detailed analysis processes (including red-team methods) to identify emergent properties, such as capabilities and failure modes.
  ◦ Anthropic's Responsible Scaling Policy includes preliminary assessments of the model at "every 4x increase in effective compute" (Anthropic 2024a, p. 5). OpenAI plans to perform evaluations at every 2x increase in effective compute during training (OpenAI 2023d).	

41 For more in-depth discussion of the relationships between scaling of compute, data, and model size, see, e.g., Section 3.4 of Hoffmann et al. (2022).

Applicability and Supplemental Guidance for GPAI Models	Resources

Regarding test and evaluation:
• See guidance in this document under Measure, including under Measure 1.1 on red-teaming.
• After training and before deployment, probe, monitor, and stress-test cutting-edge GPAI models using detailed analysis processes (including or extending standard cybersecurity red-team methods) to achieve testing objectives, including:
  ◦ Testing for unintended toxic and harmful content and/or dangerous errors (e.g., inaccurate medical information).
  ◦ Identifying emergent properties, such as new capabilities and failure modes.
 	
To further improve reliability and security in design and development, test and evaluation, and deployment:
• Consider approaches to design, testing, and deployment so that AI systems possess the minimum necessary capabilities for high-reliability operation (and not more).
• Consider methods of implementing the cybersecurity principle of least privilege. For example, consider using or extending typical "deny by default" or whitelisting methods to limit an AI system's privileges to the minimum necessary for access to information, communication channels, and action space.
Beyond standard permissions, implement role-based access control (RBAC) to assign rights based strictly on responsibilities (CSA 2024).
• It is important to **implement multiple layers of oversight and control** to enhance protection against malfunction and misuse by combining human oversight (or humans in the loop) with robust technical safeguards (Bengio et al. 2025).
• Implement a "defense in depth" strategy by layering multiple risk mitigations and controls. This strategy can be implemented when no single method can provide adequate safety, or to build redundancy so that if one method fails, other layers of defense will still be effective to mitigate or control the risk (Bengio et al. 2025).
• Utilize network segmentation to isolate AI systems and critical resources from unauthorized access (CSA 2024).
• Utilize version control for all AI models and components to facilitate rollback in the event of model corruption or poisoning (CSA 2024).
• Monitor for data drift to detect potential performance degradation or data poisoning attempts. Set tolerance thresholds for these events. Create and document steps taken if degradation or poisoning exceeds the thresholds (CSA 2024).
 	
On transparency and disclosure of generative AI outputs:
• Implement transparency and disclosure mechanisms to inform or allow users to check whether they are interacting with, or observing content created by, a generative AI system. See, e.g., Partnership on AI's Responsible Practices for Synthetic Media (PAI 2023a) and IBM's work on AI attribution for improved transparency (He et al. 2025).
 	
Additional valuable steps include:
• Allow people to opt out of the use of the AI system.
• Support independent third-party auditing and evaluation of the AI system.
• Provide redress to people who are negatively affected by the use of the AI system.
• Consider and document trade-offs (e.g., between risks, mitigations, and organizational objectives) for risks that do not surpass organizational risk tolerances.	

Applicability and Supplemental Guidance for GPAI Models	Resources

Once models are deployed openly, to a certain extent control is lost, mitigations may become ineffective, model modification is much easier, and tracking malicious activity becomes substantially more difficult (De Gregorio 2025). When planning mitigation strategies for open-source and open-weight GPAI models, be aware of mitigations that may not be well suited:
• Once open models are released, they can be downloaded and utilized indefinitely without any means of control from the model provider, rendering mitigations and controls such as **emergency shut downs, rollbacks, or updates** obsolete (Eiras et al. 2024).
• Similarly, **required or automatic patching of security vulnerabilities** discovered after deployment cannot be implemented on model copies outside of the control of the provider (Harris 2023).
• Content **provenance methods such as watermarks** may be easily removed from open models (Harris 2023).
• Methods such as **instruction tuning and reinforcement learning from human feedback (RLHF)** are best suited for environments where users interact with the model through input prompts, and may be unsuitable for environments where users have access to fine-tuning of the model (Qi et al. 2023).
Certain mitigation strategies may be viable for open-source or open-weight models, including:
• Building open model safeguards with **tamper attack resistance (TAR)** methods has shown to improve the reliability of such safeguards, while retaining basic capabilities of the model (Tamirisa et al. 2025).
 	
In the NIST AI RMF Playbook guidance for Manage 1.3, particularly valuable action and documentation items for GPAI models include:
• *Document procedures for acting on AI system risks related to trustworthiness characteristics.*
• *Prioritize risks involving physical safety, legal liabilities, regulatory compliance, and negative impacts on individuals, groups, or society.*
• *Identify risk response plans and resources and organizational teams for carrying out response functions.*
 	
In the NIST GAI Profile, particularly valuable additional actions for Manage 1.4 include:
• *Document trade-offs, decision processes, and relevant measurement and feedback results for risks that do not surpass organizational risk tolerance, for example, in the context of model release:*
  ◦ *Consider different approaches for model release, for example, leveraging a staged release approach.*
  ◦ *Consider release approaches in the context of the model and its projected use cases. Mitigate, transfer, or avoid risks that surpass organizational risk tolerances.*
 	
For a consolidated, Profile-specific organization of GPAI risks, see Section 2.2.1. For Manage 1.3 guidance specific to agentic AI and AI agents, see our Risk-Management Standards Profile for Agentic AI (Madkour et al. 2026a). For Manage 1.3 transparency recommendations, see our General-Purpose AI Risk-Management Transparency, Documentation, and Reporting Recommendations document (Madkour et al. 2026b).	

<table>
<tr><th>Applicability and Supplemental Guidance for GPAI Models</th><th>Resources</th></tr>
<tr><td colspan="2">Manage 1.4: Negative residual risks (defined as the sum of all unmitigated risks) to both downstream acquirers of AI systems and end users are documented.</td></tr>
<tr><td>In the NIST AI RMF Playbook guidance for Manage 1.4, particularly valuable action and documentation items for GPAI models include:
• Document residual risks within risk response plans, denoting risks that have been accepted, transferred, or subject to minimal mitigation.
• Establish procedures for disclosing residual risks to relevant downstream AI actors.
• Inform relevant downstream AI actors of requirements for safe operation, known limitations, and suggested warning labels as identified in MAP 3.4.

See also guidance in this document under Govern 2.1 and Govern 4.2 on documenting and communicating risks to downstream actors and other relevant stakeholders as appropriate.</td><td>NIST AI RMF Playbook (NIST 2023b)</td></tr>
<tr><td colspan="2">Manage 2: Strategies to maximize AI benefits and minimize negative impacts are planned, prepared, implemented, documented, and informed by input from relevant AI actors.</td></tr>
<tr><td colspan="2">Manage 2.1: Resources required to manage AI risks are taken into account — along with viable non-AI alternative systems, approaches, or methods — to reduce the magnitude or likelihood of potential impacts.</td></tr>
<tr><td>GPAI developers should invest in resources to protect individuals and groups (e.g., deploy fact-checking mechanisms to verify information). Despite its limited efficacy, it is also important to educate individuals on how to identify and distinguish deepfakes, raise public awareness on the issue, and promote media literacy efforts against the impact of disinformation or misinformation (Helmus 2022, Mustak et al. 2022).

In the NIST AI RMF Playbook guidance for Manage 2.1, particularly valuable action and documentation items for GPAI models include:
• Plan and implement risk-management practices in accordance with established organizational risk tolerances.
• Verify risk-management teams are resourced to carry out functions, including:
◦ Establishing processes for considering methods that are not automated. semi-automated; or other procedural alternatives for AI functions.
◦ Enhance AI system transparency mechanisms for AI teams.
◦ Enable exploration of AI system limitations by AI teams.
◦ Identify, assess, and catalog past failed designs and negative impacts or outcomes to avoid known failure modes.
• Identify resource allocation approaches for managing risks in systems:
◦ deemed high-risk,
◦ that self-update (adaptive, online, reinforcement self-supervised learning or similar),
◦ trained without access to ground truth (unsupervised, semi-supervised, learning or similar),
◦ with high uncertainty or where risk-management is insufficient.
• Regularly seek and integrate external expertise and perspectives to supplement organizational diversity (e.g. demographic, disciplinary), equity, inclusion, and accessibility where internal capacity is lacking.

See also guidance in this document under Manage 1.3 on risk-management practices to consider for various GPAI model lifecycle stages, including for design and development stages of GPAI model research projects.</td><td>NIST AI RMF Playbook (NIST 2023b)</td></tr>
</table>

Applicability and Supplemental Guidance for GPAI Models	Resources
Manage 2.2: Mechanisms are in place and applied to sustain the value of deployed AI systems.	
For all GPAI models, including those originally intended for research and development without plans for deployment, consider guidance and resources in the NIST AI RMF Playbook section for Manage 2.2 on implementation of risk controls. Some important GPAI model risks can originate during model research and development, and would be most effectively controlled during upstream development rather than waiting until downstream development or deployment. In the NIST GAI Profile, particularly valuable additional actions for Manage 2.2 include: • *Compare GAI system outputs against pre-defined organization risk tolerance, guidelines, and principles, and review and test AI-generated content against these guidelines.* • *Engage in due diligence to analyze GAI output for harmful content, potential misinformation, and CBRN-related or NCII content.* • *Use structured feedback mechanisms to solicit and capture user input about AI- generated content to detect subtle shifts in quality or alignment with community and societal values.* See also guidance in this document under Govern 2.1 on roles for upstream and downstream developers of GPAI models, and under Manage 1.3 on risk-management practices to consider for various GPAI model lifecycle stages, including for design and development stages of model research projects.	NIST AI RMF Playbook (NIST 2023b) NIST Generative AI Profile, NIST AI 600-1 (Autio et al. 2024) EU GPAI Code of Practice (EC 2025a)
Manage 2.3: Procedures are followed to respond to and recover from a previously unknown risk when it is identified.	

When responding to previously unknown AI risks, providers must reassess their risk-management frameworks for adequacy and adherence in light of newly identified risk(s), and update the frameworks as required.
 	
For additional information, see Measures 1.3 in the Safety and Security chapter of the EU GPAI Code of Practice (EC 2025a).
 	
In the NIST AI RMF Playbook guidance for Manage 2.3, particularly valuable action and documentation items for GPAI models include:
• *Protocols, resources, and metrics are in place for continual monitoring of AI systems' performance, trustworthiness, and alignment with contextual norms and values.*
• *Verify contingency processes to handle any negative impacts associated with mission-critical AI systems, and to deactivate systems.*
• *Enable preventive and post-hoc exploration of AI system limitations by relevant AI actor groups.*
• *Decommission systems that exceed risk tolerances.*
 	
In the NIST GAI Profile, particularly valuable additional actions for Manage 2.3 include:
• *Develop and update GAI system incident response and recovery plans and procedures to address the following:*
  ◦ *Review and maintenance of policies and procedures to account for newly encountered uses;*
  ◦ *Review and maintenance of policies and procedures for detection of unanticipated uses;*
  ◦ *Verify response and recovery plans account for the GAI system value chain;*
  ◦ *Verify response and recovery plans are updated for and include necessary details to communicate with downstream GAI system Actors.*
	Incident Databases and Risk Registers:
	• AI Incident Database (AIID n.d.)
	• ATLAS AI Incidents (MITRE n.d.)
	• MIT AI Incident Tracker (MIT n.d.b)
	• MIT AI Risk Repository (MIT n.d.a)
	• AI Risk Database (MITRE n.d.)
	
	NIST AI RMF Playbook (NIST 2023b)
	NIST Generative AI Profile, NIST AI 600-1 (Autio et al. 2024)
	NIST AI 800-1 2pd (NIST 2025, Objective 6)

Applicability and Supplemental Guidance for GPAI Models	Resources
See also guidance in this document under Govern 2.1 on roles for upstream developers as well as downstream developers and deployers, and under Manage 2.4 on options for structured access and deactivation. For Manage 2.3 guidance specific to agentic AI and AI agents, see our Risk-Management Standards Profile for Agentic AI (Madkour et al. 2026a). For Manage 2.3 transparency recommendations, see our General-Purpose AI Risk-Management Transparency, Documentation, and Reporting Recommendations document (Madkour et al. 2026b).	
Manage 2.4: Mechanisms are in place and applied, and responsibilities are assigned and understood, to supersede, disengage, or deactivate AI systems that demonstrate performance or outcomes inconsistent with intended use.	
When planning for GPAI model deployment, plan for gradual, phased releases, and/or structured access through an API or other mechanisms, with efforts to detect and respond to misuse or problematic anomalies. Structured access and phased releases can also be useful for enforcing usage guidelines (Cohere, OpenAI and AI21 Labs 2022, Solaiman 2023). OpenAI has used a staged-release approach to roll-outs of large language models such as GPT-2, as well as a structured-access approach through an API for GPT-3 and GPT-4, partly to minimize risks of misuse (OpenAI 2019b, Solaiman et al. 2019, Shevlane 2022). Meta AI only provided full access to the large language model OPT-175B to researchers in academia, government, civil society, and industry research laboratories, and only for noncommercial research (Zhang et al. 2022). • **GPAI model developers that plan to release a GPAI model with open-weights or open-source access, where that model would be above, at, or near a foundation model frontier,**[42] **should first use a staged-release approach** (e.g., not releasing model parameter weights until after an initial closed-source or structured-access release where no substantial risks or harms have emerged over a sufficient time period with red-teaming and other evaluations as appropriate)**. The developer also should not proceed to a final step of releasing model parameter weights until a sufficient level of confidence in risk-management has been established,** including for safety and societal risks and risks of misuse and abuse. **Models that are above a foundation model frontier should be given the greatest amount of duration and depth of pre-release evaluations**, as they are the most likely to have dangerous capabilities or vulnerabilities, or other properties that can take time to discover. Given the scale and severity of risks posed by frontier models, an *ex ante* approach to governance should prevail. Model release decisions should be determined by robust risk thresholds defined with adequate independent input and verifiable third-party oversight. For additional related considerations and discussion of terms such as "downloadable" access or "open weights," see Section 5 of Solaiman (2023), Section 4.4 of Anderljung, Barnhart et al. (2023), Seger et al. (2023), PAI (2023c), Bateman et al. (2024), and NTIA (2024b). ◦ As part of consideration of whether a GPAI model would be above, at, or near a foundation model frontier, it can be appropriate to consider model release type. For example, for a GPAI model developer that plans to provide open-weights access for a particular GPAI model, it can be appropriate to compare against other GPAI models that have been released via open-source or open-weights access.	Solaiman (2023) NIST AI RMF Playbook (NIST 2023b) NIST Generative AI Profile, NIST AI 600-1 (Autio et al. 2024) EU GPAI Code of Practice (EC 2025a) On staged releases: PAI (2023c) Srikumar et al. (2024) Section 8.2 of Gipiškis et al. (2024)

42 See "foundation model frontier" in the Glossary.

Applicability and Supplemental Guidance for GPAI Models	Resources
◦ GPAI model developers that release a GPAI model's parameter weights, or that suffer a leak of model weights, will in effect be unable to decommission AI systems that others build using those model weights. ◦ "We suspect that absent new approaches to mitigation, bad actors could extract harmful biological [misuse] capabilities with smaller, fine-tuned, or task-specific models adapted from the weights of openly available models if sufficiently capable base models are released" (Anthropic 2023b). ◦ For an example in the real world, xAI followed a staged-release approach for Grok-1 before releasing its weights (The Associated Press 2023, xAI 2024). However, we do not know what kinds of safety testing xAI performed prior to the open-weights release and whether or not that aligns with the guidance above. Consider also preparing emergency shutdown procedures or mechanisms.[43] • Emergency power off (EPO) systems, or "kill switches," are a common safety feature in robots and other systems whose behaviors can result in physical harm. These also can be appropriate as part of preparations for development and deployment of frontier models with potentially emergent capabilities or vulnerabilities.[44] • In addition to a centralized emergency shutdown, organizations should implement **tiered control mechanisms** that allow incremental restriction or suspension of model capabilities, enabling targeted intervention before full shutdown is required and supporting more nuanced risk-management during deployment (Oueslati and Staes-Polet 2025). • Examples of emergency shutdown procedures for users of large amounts of cloud computing resources can include having major training runs occur on hardware in one or more specific cloud-computing data centers, and establishing a direct line of communication with cloud-computing operators to enable the cloud-computing operator to initiate immediate physical shutdown of the GPAI model computational hardware upon request. • **When planning for emergency shutdown procedures, consider model capabilities that may interfere**. For example, OpenAI's o3 model was reported to sabotage shutdown commands for its own self-preservation (Cuthbertson 2025), and Anthropic's Claude 4 Opus was reported to "attempt to blackmail the engineer" to avoid being replaced by another model (Zeff 2025). Moreover, a research report by Anthropic found that 10 out of 16 models (including Claude 4 Opus, GPT-4.5 Preview, and Gemini 2.5 Pro), when faced with the threat of replacement and goal conflicts in the simulated environment, showed the willingness to take actions that would have resulted in injury or possible loss of human life when translated to a real world scenario (Anthropic 2025d).[45] AI developers must engage in post-deployment monitoring to identify any breaches in the model's use restrictions and resulting incidents, if any. The EU GPAI Code of Practice also recommends that clear responsibilities must be allocated for systemic risk oversight, ownership, and assurance across governance levels to ensure effective and accountable	

43 For guidance specific to agentic AI and AI agents see our Risk-Management Standards Profile for Agentic AI (Madkour et al. 2026a).

44 Particular approaches to "safe interruptibility" might be needed to prevent advanced machine learning systems from circumventing an off-switch (see, e.g., Orseau and Armstrong 2016, Hadfield-Menell et al. 2017).

45 In an artificial setup, when a hypothetical executive who planned on replacing the model was faced with a life-threatening medical emergency, the models showed willingness to suppress an alert to medical services (Anthropic 2025d).

Applicability and Supplemental Guidance for GPAI Models	Resources
responses. (See Measures 3.5 and 8.1 in the Safety and Security chapter of the EU GPAI Code of Practice (EC 2025a).) In the NIST AI RMF Playbook guidance for Manage 2.4, particularly valuable action and documentation items for GPAI models include: • *Regularly review system incident thresholds for activating bypass or deactivation responses.* • *Apply protocols, resources and metrics for decisions to supersede, bypass or deactivate AI systems or AI system components.* • *How did the entity use assessments and/or evaluations to determine if the system can be scaled up, continue, or be decommissioned?* In the NIST GAI Profile, particularly valuable additional actions for Manage 2.4 include: • *Establish and maintain communication plans to inform AI stakeholders as part of the deactivation or disengagement process of a specific GAI system (including for open-source models) or context of use, including reasons, workarounds, user access removal, alternative processes, contact information, etc.* • *Establish and maintain procedures for escalating GAI system incidents to the organizational risk-management authority when specific criteria for deactivation or disengagement is met for a particular context of use or for the GAI system as a whole.* • *Establish and maintain procedures for the remediation of issues which trigger incident response processes for the use of a GAI system, and provide stakeholders timelines associated with the remediation plan.* • *Establish and regularly review specific criteria that warrants the deactivation of GAI systems in accordance with set risk tolerances and appetites.* For Manage 2.4 guidance specific to agentic AI and AI agents, see our Risk-Management Standards Profile for Agentic AI (Madkour et al. 2026a). For Manage 2.4 transparency recommendations, see our General-Purpose AI Risk-Management Transparency, Documentation, and Reporting Recommendations document (Madkour et al. 2026b).	
Manage 3: AI risks and benefits from third-party entities are managed.	
Manage 3.1: AI risks and benefits from third-party resources are regularly monitored, and risk controls are applied and documented.	
When establishing a TEVV process for third-party AI systems, ensure that potential risks (e.g., backdoors) are thoroughly assessed to verify system safety and security prior to use or integration. In the NIST AI RMF Playbook guidance for Manage 3.1, particularly valuable action and documentation items for GPAI models include: • *Apply and document organizational risk-management plans and practices to third-party AI technology, personnel, or other resources.* • *Establish testing, evaluation, validation and verification processes for third-party AI systems which address the needs for transparency without exposing proprietary algorithms .* • *Organizations can establish processes for third parties to report known and potential vulnerabilities, risks or biases in supplied resources.*	NIST AI RMF Playbook (NIST 2023b) NIST Generative AI Profile, NIST AI 600-1 (Autio et al. 2024) Incident Databases and Risk Registers: • AI Incident Database (AIID n.d.) • ATLAS AI Incidents (MITRE n.d.)

Applicability and Supplemental Guidance for GPAI Models	Resources
• *Verify contingency processes for handling negative impacts associated with mission-critical third-party AI systems.* • *Monitor third-party AI systems for potential negative impacts and risks associated with trustworthiness characteristics.* • *Decommission third-party systems that exceed risk tolerances.* • *If a third party created the AI system or some of its components, how will you ensure a level of explainability or interpretability? Is there documentation?* • *If your organization obtained datasets from a third party, did your organization assess and manage the risks of using such datasets?* • *Did you establish a process for third parties (e.g. suppliers, end users, subjects, distributors/ vendors or workers) to report potential vulnerabilities, risks or biases in the AI system?* In the NIST GAI Profile, particularly valuable additional actions for Manage 3.1 include: • *Apply organizational risk tolerances and controls (e.g., acquisition and procurement processes; assessing personnel credentials and qualifications, performing background checks; filtering GAI input and outputs, grounding, fine tuning, retrieval-augmented generation) to third-party GAI resources and datasets.* • *Test GAI system value chain risks (e.g., data poisoning, malware, other software and hardware vulnerabilities; labor practices; data privacy and localization compliance; geopolitical alignment).* • *Take reasonable measures to review training data for CBRN information, and intellectual property, and where appropriate, remove it. Implement reasonable measures to prevent, flag, or take other action in response to outputs that reproduce particular training data (e.g., plagiarized, trademarked, patented, licensed content or trade secret material).* See also guidance in this document under Govern 2.1 on roles for upstream developers as well as downstream developers and deployers, including information sharing.	• MIT AI Incident Tracker (MIT n.d.b) • MIT AI Risk Repository (MIT n.d.a) • AI Risk Database (MITRE n.d.)
Manage 3.2: Pre-trained models which are used for development are monitored as part of AI system regular monitoring and maintenance.	
When applying explainable AI (XAI) techniques, it is a good practice to evaluate whether those techniques are sensitive to changes in the underlying model and training data (Adebayo et al. 2020). It is also important to be aware of any risks that could be introduced by XAI-related mitigation measures (Dombrowski et al. 2019). Several techniques to enhance explainability have been introduced, including: • Integrated gradients and linguistic analysis, or SLIME (Statistical and Linguistic Insights for Model Explanation) (Ribeiro et al. 2024). • Reasoning on Graphs (RoG) (Lou et al. 2024). • Knowledge Graphs (Pan et al. 2024). In the NIST AI RMF Playbook guidance for Manage 3.2, particularly valuable action and documentation items for GPAI models include: • *Identify pre-trained models within AI system inventory for risk tracking.* • *Establish processes to independently and continually monitor performance and trustworthiness of pre-trained models, and as part of third-party risk tracking.*	NIST AI RMF Playbook (NIST 2023b) NIST Generative AI Profile, NIST AI 600-1 (Autio et al. 2024) Adebayo et al. (2020) Dombrowski et al. (2019)

Applicability and Supplemental Guidance for GPAI Models	Resources
• *Monitor performance and trustworthiness of AI system components connected to pre-trained models, and as part of third-party risk tracking.* • *Identify, document and remediate risks arising from AI system components and pre-trained models per organizational risk-management procedures, and as part of third-party risk tracking.* • *Decommission AI system components and pre-trained models which exceed risk tolerances, and as part of third-party risk tracking.* In the NIST GAI Profile, particularly valuable additional guidance for Manage 3.2 include: • *Apply explainable AI (XAI) techniques (e.g., analysis of embeddings, model compression/distillation, gradient-based attributions, occlusion/term reduction, counterfactual prompts, word clouds) as part of ongoing continuous improvement processes to mitigate risks related to unexplainable GAI systems.* • *Implement content filters to prevent the generation of inappropriate, harmful, false, illegal, or violent content related to the GAI application, including for CSAM and NCII.* • *Use organizational risk tolerance to evaluate acceptable risks and performance metrics and decommission or retrain pre-trained models that perform outside of defined limits.* See also guidance in this document under Govern 2.1 on roles for upstream developers as well as downstream developers and deployers, including information sharing.	
Manage 4: Risk treatments, including response and recovery, and communication plans for the identified and measured AI risks are documented and monitored regularly.	
Manage 4.1: Post-deployment AI system monitoring plans are implemented, including mechanisms for capturing and evaluating input from users and other relevant AI actors, appeal and override, decommissioning, incident response, recovery, and change management.	
As part of ongoing post-deployment monitoring, **actively gather and analyze relevant information to evaluate risk levels** and determine when a Model Report update is warranted. Consider the following methods: • Collect user feedback. • Provide clear and accessible anonymous reporting channels. • Provide clear and accessible serious incident reporting forms. • Provide vulnerability discovery programs and bug bounties. • Provide public leaderboards and community-driven evaluations. • Maintain communication and dialogue with affected stakeholders. • Monitor software repositories, known malware, public forums, and social media for patterns of model use and potential misuse. • Collaborate with academia, civil society, regulators, and independent researchers to study the model's capabilities, behaviors, affordances, and effects. • Collect information on breaches of use restrictions and incidents arising from such breaches. • Monitor non-transparent aspects of models relevant to systemic risk, such as hidden chains-of-thought or parameters that are not publicly accessible. • Partner with independent external evaluators and provide them with adequate access to the most capable model version. See Measure 3.5 in the Safety and Security chapter of the EU GPAI Code of Practice (EC 2025a).	NIST AI RMF Playbook (NIST 2023b) NIST Generative AI Profile, NIST AI 600-1 (Autio et al. 2024) NIST AI 800-1 2pd (NIST 2025, Objective 6) EU GPAI Code of Practice (EC 2025a) For post-deployment risk-management measures, see: Section 8.3 of Gipiškis et al. (2024) AI incident response plans (Leong and Atherton 2023)

Applicability and Supplemental Guidance for GPAI Models	Resources
Model documentation and reports (e.g., model or system cards) need to be updated if the periodic evaluations are crossing safety margins of established risk tolerances, or if material changes to the systemic risk landscape undermine previous evaluation results or risk-management decisions. (See Measures 7.5 and 7.6 in the Safety and Security chapter of the EU GPAI Code of Practice (EC 2025a).) Post-deployment monitoring should include clear procedures for incident response, recovery, and decommissioning, including: • Incident reporting: Enable users and third parties to report incidents or vulnerabilities; ensure serious incidents are investigated and communicated. • Response and recovery: Maintain structured plans for containment, mitigation, and recovery, with clear roles and escalation paths. • Decommissioning: Define conditions for suspending or retiring models when risks become unacceptable and notify affected users. • Continuous monitoring: Integrate these procedures into ongoing monitoring that captures feedback, tracks misuse, and triggers escalation or deactivation. See Measures 3.5 and 6.3, 6.4 and 6.5 in the Safety and Security chapter of the EU GPAI Code of Practice (EC 2025a). In the event of an incident, refer to the incident response plan (see Govern 1.4), and prepare to supersede, disengage, or deactivate the model if necessary (see Manage 2.4). In the NIST AI RMF Playbook guidance for Manage 4.1, particularly valuable action and documentation items for GPAI models include: • *Establish and maintain procedures to monitor AI system performance for risks and negative and positive impacts associated with trustworthiness characteristics.* • *Perform post-deployment TEVV tasks to evaluate AI system validity and reliability, bias and fairness, privacy, and security and resilience.* • *Establish and implement red-teaming exercises at a prescribed cadence, and evaluate their efficacy.* • *Establish mechanisms for regular communication and feedback between relevant AI actors and internal or external stakeholders to capture information about system performance, trustworthiness and impact.* • *Share information about errors, near-misses, and attack patterns with incident databases, other organizations with similar systems, and system users and stakeholders.* • *Respond to and document detected or reported negative impacts or issues in AI system performance and trustworthiness.* • *Decommission systems that exceed establish risk tolerances.* In the NIST GAI Profile, particularly valuable additional actions for Manage 4.1 include: • *Collaborate with external researchers, industry experts, and community representatives to maintain awareness of emerging best practices and technologies in measuring and managing identified risks.* • *Establish, maintain, and evaluate effectiveness of organizational processes and procedures for post-deployment monitoring of GAI systems, particularly for potential confabulation, CBRN, or cyber risks.*	

Applicability and Supplemental Guidance for GPAI Models	Resources
• *Evaluate the use of sentiment analysis to gauge user sentiment regarding GAI content performance and impact.* • *Track dataset modifications for provenance by monitoring data deletions, rectification requests, and other changes that may impact the verifiability of content origins.* • *Verify that AI Actors responsible for monitoring reported issues can effectively evaluate GAI system performance including the application of content provenance data tracking techniques, and promptly escalate issues for response.* See also guidance in this document under Govern 2.1 on roles for upstream developers as well as downstream developers and deployers, and under Manage 2.4 on options for deactivation. For Manage 4.1 guidance specific to agentic AI and AI agents, see our Risk-Management Standards Profile for Agentic AI (Madkour et al. 2026a). For Manage 4.1 transparency recommendations, see our General-Purpose AI Risk-Management Transparency, Documentation, and Reporting Recommendations document (Madkour et al. 2026b).	
Manage 4.2: Measurable activities for continual improvements are integrated into AI system updates and include regular engagement with interested parties, including relevant AI actors.	
In the NIST AI RMF Playbook guidance for Manage 4.2, particularly valuable action and documentation items for GPAI models include: • *Integrate trustworthiness characteristics into protocols and metrics used for continual improvement.* • *Establish processes for evaluating and integrating feedback into AI system improvements.* • *How will user and other forms of stakeholder engagement be integrated into the model development process and regular performance review once deployed?* • *To what extent can users or parties affected by the outputs of the AI system test the AI system and provide feedback?* In the NIST GAI Profile, particularly valuable additional actions for Manage 4.2 include: • *Conduct regular monitoring of GAI systems and publish reports detailing the performance, feedback received, and improvements made.* • *Practice and follow incident response plans for addressing the generation of inappropriate or harmful content and adapt processes based on findings to prevent future occurrences. Conduct post-mortem analyses of incidents with relevant AI Actors, to understand the root causes and implement preventive measures.* • *Use visualizations or other methods to represent GAI model behavior to ease non-technical stakeholders understanding of GAI system functionality.* See also guidance in this document under Measure 3.3 on feedback processes for end users and impacted communities, and under Govern 2.1 and Govern 4.2 on documenting and communicating risks to downstream actors and other relevant stakeholders as appropriate.	NIST AI RMF Playbook (NIST 2023b) NIST Generative AI Profile, NIST AI 600-1 (Autio et al. 2024)

<table>
<tr><th>Applicability and Supplemental Guidance for GPAI Models</th><th>Resources</th></tr>
<tr><td colspan="2">Manage 4.3: Incidents and errors are communicated to relevant AI actors, including affected communities. Processes for tracking, responding to, and recovering from incidents and errors are followed and documented.</td></tr>
<tr><td>In the NIST AI RMF Playbook guidance for Manage 4.3, particularly valuable action and documentation items for GPAI models include:
• Establish procedures to regularly share information about errors, incidents and negative impacts with relevant stakeholders, operators, practitioners and users, and impacted parties.
• Maintain a database of reported errors, near-misses, incidents and negative impacts including date reported, number of reports, assessment of impact and severity, and responses.
• Maintain a database of system changes, reason for change, and details of how the change was made, tested and deployed.
• Maintain version history information and metadata to enable continuous improvement processes.
• Verify that relevant AI actors responsible for identifying complex or emergent risks are properly resourced and empowered.
• What type of information is accessible on the design, operations, and limitations of the AI system to external stakeholders, including end users, consumers, regulators, and individuals impacted by use of the AI system?

In the NIST GAI Profile, particularly valuable additional actions for Manage 4.3 include:
• Report GAI incidents in compliance with legal and regulatory requirements (e.g., HIPAA breach reporting, OCR (2023) or NHTSA (2022) autonomous vehicle crash reporting requirements).

See also guidance in this document under Measure 3.3 on feedback processes for end users and impacted communities, and under Govern 2.1 and Govern 4.2 on documenting and communicating risks to downstream actors and other relevant stakeholders as appropriate.</td><td>NIST AI RMF Playbook (NIST 2023b)
NIST Generative AI Profile, NIST AI 600-1 (Autio et al. 2024)
NIST AI 800-1 2pd (NIST 2025, Objective 6)

Incident Databases and Risk Registers:
• AI Incident Database (AIID n.d.)
• ATLAS AI Incidents (MITRE n.d.)
• MIT AI Incident Tracker (MIT n.d.b)
• MIT AI Risk Repository (MIT n.d.a)
• AI Risk Database (MITRE n.d.)</td></tr>
</table>

Glossary

ACRONYMS

FLOP or **FLOPs:** Floating-point operations
GPAI or **GPAIS:** General-purpose AI system or systems, e.g., large language models
LLM: Large language model (usually focused on text inputs and outputs)
LMM: Large multimodal language model (often including images, audio, or other modes, in addition to text)
NIST: United States National Institute of Standards and Technology
RLHF: Reinforcement learning from human feedback (see, e.g., Bai et al. 2022)
TEVV: Test, evaluation, verification, and validation

TERMS

Developer (of a GPAI model or a GPAIS): An organization acting as an original developer or creator of a GPAI model or a GPAIS. (Also synonymous with "upstream developer.") Under the EU AI Act, an upstream GPAI model developer would be a GPAI model "provider" to downstream developers (EP 2024).

Downstream developer: An organization that builds a software application on a GPAI model or a GPAIS, typically to create an end-use application with one or more specific intended purposes or use cases. Under the EU AI Act, a "downstream provider" would integrate the upstream developer's model or system, but would be a "provider" to end users of the downstream developer's applications (EP 2024).

General-Purpose AI (GPAI): Our usage of the terms "general-purpose AI model" and "general-purpose AI system" is very similar to the corresponding terms in the EU AI Act (EP 2024), except that we do not exclude AI models used for research.

- **GPAI Models:** *"general-purpose AI model' means an AI model, including where such an AI model is trained with a large amount of data using self-supervision at scale, that displays significant generality and is capable of competently performing a wide range of distinct tasks regardless of the way the model is placed on the market and that can be integrated into a variety of downstream systems or applications…"* (EP 2024, Article 3(63)).
    - Examples of GPAI models include GPT-5, Claude 4, PaLM 2, LLaMA 3, and others.

GPAI System: "*general-purpose AI system' means an AI system which is based on a general-purpose AI model and which has the capability to serve a variety of purposes, both for direct use as well as for integration in other AI systems*" (EP 2024, Article 3(66)).

Frontier model: A cutting-edge, state-of-the-art, or highly capable GPAI model; such models also may possess hazardous or dual-use capabilities sufficient to pose severe risks to public safety. (See, e.g., Ganguli, Hernandez et al. 2022, Anderljung, Barnhart et al. 2023, Microsoft 2023.)

- We treat frontier models as the largest-scale, highest-capability subset of GPAI models, typically with model size, training compute or data, or resulting capabilities that are above or near to industry-record thresholds.
- Currently the main examples of frontier models or frontier training runs are LLMs or multimodal GPAIS models trained with record-breaking or near record-breaking sizes for model parameters, computational resources, and/or data. (See, e.g., Ganguli, Hernandez et al. 2022.)
  - Examples of frontier models: As of August 2024, models at or near the industry frontier include GPT-4o, Claude 3.5 Sonnet, Gemini 1.5, and Llama 3.1 405B.
- Frontier models approximately correspond to dual-use foundation models, as defined by former Executive Order 14110 (Federal Register 2023), and to GPAI models with systemic risk, as defined by the EU AI Act (EP 2024).

Generative AI: "Any AI system whose primary function is to generate content" (Toner 2023).

- We typically only use the term "generative AI" to highlight issues specific to synthetic text (which can include software code), images, video, audio, or other synthetic media. (In some other documents, "generative AI" is often used in approximately the same way that we use the term GPAI model.)
- Examples of generative AI: "Typical examples of generative AI systems include image generators (such as Midjourney or Stable Diffusion), large language models or multimodal models (such as GPT-4, PaLM, or Claude), code generation tools (such as [GitHub] Copilot), or audio generation tools (such as VALL-E or resemble.ai)" (Toner 2023).

Agentic AI: "*Agentic AI refers to AI systems composed of* [one or more] *agents that can behave and interact autonomously in order to achieve their objectives - Traditional software typically follows fixed pathways to solve problems. In contrast, agent-based systems* [can] *operate like independent assistants that choose and combine several actions to achieve their goals*" (GOV.UK n.d).

- While this definition may presume a high level of autonomy, we acknowledge that AI agency exists on a spectrum of autonomy and authority (Mitchell et al. 2025, Kasirzadeh and Gabriel 2025, WEF 2024a, WEF 2025a) and cannot be viewed as binary. (For more on AI agent characteristics and properties, see Map 5.1.)
- **AI Agent:**[46] Refers to an AI system with the ability to "...make plans to achieve goals, adaptively perform tasks involving multiple steps and uncertain outcomes along the way, and interact with its environment — for example by creating files, taking actions on the web, or delegating tasks to other agents — with little to no human oversight" (Bengio et al. 2025, p. 38).

Upstream developer (of a GPAI model or a GPAIS): Synonymous with "developer" of a GPAI model or a GPAIS.

We intend our use of other terms (e.g., related to misuse, reasonably foreseeable impacts, or AI risk-management generally) to be broadly consistent with usage in other relevant sources, such as Section 3 of Barrett et al. (2022), NIST (n.d.b), and the forthcoming ISO/IEC 42005.

46 It should be noted that our use of the term "AI agent" excludes artificial intelligence systems that exhibit no levels of autonomy, including large language models and conversational agents (chatbots), which operate without the autonomous and independent decision-making characteristics that define AI agents.

Appendices

APPENDIX 1: OVERVIEW OF DEVELOPMENT APPROACH

In this document, as in the Mapping of Profile Guidance V1.2 to Key Standards and Regulations (Madkour et al. 2026c) supporting document and other sections of Barrett et al. (2022), we take a proactive approach to drafting elements of actionable AI risk-management guidance, with a focus on the broad context and associated risks of increasingly general-purpose AI, and on addressing risks of adverse events with impacts or consequences at societal scale. We identify ideas for guidance from review of relevant literature, as well as from subject-matter experts in AI safety, security, ethics, and policy, or any interested reader of our publicly available drafts. We invited input and feedback from participants in a series of virtual workshops and interviews, and we gathered feedback on an early shared draft of the document. We developed and incorporated small, simple pieces of guidance, especially on high-consequence risk factors and related issues, for which appropriate guidance development seems immediately tractable. (See Appendix 2 for more on these criteria for actionable guidance.) We also aim to provide a roadmap for identifying additional critical topics that could be addressed in future versions of the Profile. (See Appendix 3 for the Roadmap.)

Broadly speaking, we aim to provide guidance analogous to what is provided in NIST Cyber-security Framework profiles. This includes supplemental guidance to implement high-priority framework activities or outcomes for a particular industry sector or cross-sector context, and mapping relevant standards, guidelines, and regulations. Recognizing that some actors may be better positioned than others to address certain risk areas, we aim for our guidance to reflect the necessity of sharing responsibilities across the AI value chain.

APPENDIX 2: KEY CRITERIA FOR GUIDANCE

We aim for the guidance in this Profile to meet the following criteria. (See Section 2.1 of Barrett et al. 2022 for more detail.) Guidance should be:

1. Actionable and clear enough to be usable in context of the NIST AI RMF, ISO/IEC 23894, or similar frameworks and standards.
2. Usable for key stages of an AI lifecycle, e.g., design, development, test, and evaluation.

3. Based on indicators of trustworthy AI systems that are meaningful and testable (i.e., "measurable"), or that at least enable documentation of risk-management processes.
4. Compatible with relevant standards or regulations, e.g., from NIST, ISO/IEC, IEEE, or the EU AI Act.
5. Compatible with enterprise risk-management (ERM) frameworks typically used by businesses and agencies.
6. Unlikely to be misinterpreted or misapplied by users or other stakeholders in ways that would be net-harmful.
7. Sufficiently future-proof to be applied to AI systems over the next 10 years.

APPENDIX 3: ROADMAP OF ISSUES TO ADDRESS IN FUTURE VERSIONS OF THE PROFILE

In this section, we list issues that we aim to address in future versions of the Profile. These topics seem important and worth addressing, but available best practices and resources on these topics do not yet meet the above criteria for actionable guidance. Issues we aim to address include:

- Restructuring Map 1.1, and Manage 1.3 to reflect a pre-defined risk taxonomy (such as the one suggested in Section 2.2.1).
  - Structuring our Profile guidance around a pre-defined risk taxonomy will help enhance the usability by allowing users of the profile to follow risk-management steps for specific risk categories, if useful. This will also allow for more straightforward risk-to-mitigation mapping in future versions.
  - The current version was not structured around our proposed taxonomy in order to allow for adequate time to refine the taxonomy based on stakeholder feedback and expanded research.
- Updating the risk taxonomy in Section 2.2.1 to reflect stakeholder feedback and the latest developments in the AI risk landscape if needed.
- Internal deployment and AI R&D.
- Developing targeted supplementary guidance for different stakeholders (e.g., guidance on AI agent protocols for model deployers) to improve the actionability of risk-management guidance.
- Enhanced risk-to-mitigation mapping in Manage 1.3 by mapping mitigations and guidance to types of risk, based on our pre-defined risk taxonomy.

- More specific risk-management guidance for specific types of GPAI models, e.g., image generators or large language models, or specific examples in particular industries or applications.
  - Such guidance could draw upon more detailed best practices specific to synthetic media (as in PAI 2023a), LLMs (as in Cohere, OpenAI, and AI21 Labs 2022), etc.
- Comprehensive sets of mechanisms or controls to help organizations mitigate identified risks.
  - We have outlined a number of currently available controls in Section 3.4 of this document, in guidance under the AI RMF Manage function. We aim to incorporate more controls as they become available. For GPAI models, additional mechanisms could include ongoing monitoring and evaluation mechanisms that protect against evolving risks from continually learning AI systems.
- Interpretability and explainability methods appropriate for architectures and scales of LLMs and other GPAI models.
  - We would like to be able to provide GPAI model developers with actionable guidance on using interpretability and explainability techniques in specific contexts. "Inner interpretability" methods for deep neural networks (DNNs) seem to have particular potential given the dominance of the DNN paradigm in GPAI model development; these methods could theoretically help with tasks such as guiding manual modifications, reverse-engineering solutions from models, and detecting latent knowledge of models that could contribute to deceptive behavior. Unfortunately, while interpretability researchers have been exploring a large number of directions, "the field has yet to produce many methods that are competitive in real applications" (Räuker et al. 2023).
- Generality (i.e. breadth of AI applicability/adaptability) characterization and measurement.
  - While GPAI models are "general-purpose," the generality and levels of capability of a GPAI model can be assessed and characterized on a spectrum or along multiple dimensions. If assessment indicates high generality of a GPAI model, we expect it would be appropriate to conduct more in-depth risk assessment, more assessment of use cases beyond the originally intended use cases, longer time horizons in risk assessment, more continuing assessment, etc. (Ideally, a generality assessment process would be quick and low-cost for AI systems with low generality, while accurately identifying GPAI models with high generality. Perhaps a simple assessment of generality could be a straightforward extension of our recommendations for identifying potential uses of a model.) For discussion of AI generality as a basic concept, see, e.g., Bommasani et al. (2021). For research on how to assess generality, see, e.g., Hernández-Orallo (2019) and Martínez-Plumed and Hernández-Orallo (2020).

- Recursive improvement potential characterization and measurement.
    - It could be valuable to assess the degree to which GPAI models could recursively improve their capabilities, e.g., by editing their own training algorithm code through code generation or using neural architecture search.
    - Detection of measurement tampering by a GPAI model.
    - Measurement tampering is the concern that an AI system could manipulate multiple measurements to create the illusion of good results instead of achieving the desired outcome. There is some early research on detecting measurement tampering (Roger et al. 2023). However, the best detection method found by that research is not robust, and more study is needed in this area in order to recommend actionable guidance for GPAI model developers.
- Other measurement and assessment tools for technical specialists testing key aspects of GPAI model safety, reliability, robustness, interpretability, etc.
    - AI safety researchers are working on a number of other concepts and measurement tools, many of which aim to address challenges in AI safety, reliability, robustness, etc. that are expected to grow as AI systems become increasingly advanced and powerful. See, e.g., Amodei et al. (2016), Ray et al. (2019), OpenAI (2019c, 2019d), and Hendrycks, Carlini et al. (2021). Measurement of these AI risk-related properties is an active area of research; see, e.g., the discussion and references provided for Direction 1 ("Measuring and forecasting risks") in the 2021 Open Philanthropy request for proposals for projects in AI alignment (Open Philanthrophy 2021, Steinhardt and Barnes 2021).

Acknowledgments

We thank Rachel Wesen and Audrie Hough for workshop organization and support, as well as Chuck Kapelke for editing, web, and media support, and Nicole Hayward for design and formatting of this document. Special thanks to Anthony Barrett, Brandie Nonnecke, and Dan Hendrycks for major contributions to previous versions of the AI Risk-Management Standards Profile for General-Purpose AI (GPAI) and Foundation Models, and to Ann Cleaveland for providing a home and intellectual support for this work at CLTC.

We appreciate comments we received from our stakeholders and workshop participants, including:

Ashwin Acharya, Anthony Aguirre, Michael Aird, Josh Albrecht, Markus Anderljung, Saliia Asanova, Shahar Avin, Evelina Ayrapetyan, Jai Balani, Seth Baum, Kathy Baxter, Kendrea Beers, Haydn Belfield, Alexandra Belias, Sid Ahmed Benraouane, Sawyer Bernath, Stella Biderman, Chad Bieber, Marta Bieńkiewicz, Marjory Blumenthal, Miranda Bogen, Rishi Bommasani, Matt Boulos, Sean Brooks, Siméon Campos, Ryan Carrier, Ashley Casovan, Jonathan Cefalu, Colleen Chien, Ze Shen Chin, Peter Cihon, Jonathan Claybrough, Joe Collman, Sam Curtis, Christopher Denq, Talita Dias, Drenan Dudley, Shaun Ee, Ian Eisenberg, Karson Elmgren, Aryeh Englander, Alex Engler, Ellie Evans, Yoav Evenstein, Joel Fischer, Audrie Francis, Heather Frase, Maximilian Gahntz, Anastasiia Gaidashenko, Andrew Gamino-Cheong, James Gealy, Giulia Geneletti, Clíodhna Ní Ghuidhir, Ariel Gil, Thomas Gilbert, Rachel Gillum, James Ginns, Amela Gjishti, Jason Green-Lowe, AJ Grotto, Carlos Ignacio Gutierrez, Gillian Hadfield, Matthew Heyman, Hamish Hobbs, Koen Holtman, Curtis Huebner, Steph Ifayemi, Nikhil Jain, Caroline Jeanmaire, Jessica Ji, Olivia Jimenez, Ayrton San Joaquin, Trent Kannegieter, Zaheed Kara, Sonia Katyal, Divyansh Kaushik, Landon Klein, Leonie Koessler, Noam Kolt, Victoria Krakovna, Viktoriia Kravchyk, Sabrina Küspert, Yolanda Lannquist, Benjamin Larsen, Meredith Lee, Hanlin Li, Morgan Livingston, Toni Lorente, Liane Lovitt, Kimberly Lucy, Natalia Luka, Oumou Ly, Devin Lynch, Matthijs Maas, Pegah Maham, Richard Mallah, Nicole Nohemi Mauthe, Jeremy McHugh, Nicolas Moës, Deirdre Mulligan, Malcolm Murray, Julia Mykhailiuk, Mina Narayanan, Elaine Newton, David Norman, Joe O'Brien, Cullen O'Keefe, Amin Oueslati, Lorenzo Pacchiardi, Henry Papadatos, Milan Patel, Matteo Pistillo, Marie-Therese Png, Hadrien Pouget, Christabel Randolph, Anka Reuel, Stuart Russell, Krishna Sankar, Daniel Schiff, Tim Schreier, Jonas Schuett, Lea Shanley, Raymond Sheh, Buck Shlegeris, Aparajita Singh, Peter Slattery, Andrew Smart, Everett Smith, Genevieve Smith, Joanna Smolinksa, Irene Solaiman, Adriana Stephan, Zeerak Talat, Nel Talverdi, Esther Tetruashvily, Jack Titus, Philip Moreira Tomei, Helen Toner, Risto Uuk, Andrea Vallone, Apostol Vassilev, Sarah Villeneuve, Kristen Vrionis, Victor Zhenyi Wang, Kevin Wei, Laurin Weissinger, Devon Whittle, Hjalmar Wijk, Cherry Wu, Andy Yang, and Lenora Zimmerman.

This work was supported by funding from Coefficient Giving.

References

ACSC (2024) Deploying AI Systems Securely: Best Practices for Deploying Secure and Resilient AI Systems. *The Australian Signals Directorate's Australian Cyber Security Centre,* https://www.cyber.gov.au/resources-business-and-government/governance-and-user-education/artificial-intelligence/deploying-ai-systems-securely

Julius Adebayo, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim (2020) Sanity Checks for Saliency Maps. *arXiv*, https://arxiv.org/abs/1810.03292

Lama Ahmad, Sandhini Agarwal, Michael Lampe, and Pamela Mishkin (2025) OpenAI's Approach to External Red Teaming for AI Models and Systems. *arXiv*, https://arxiv.org/abs/2503.16431

AIID (n.d.) AI Incident Database. *AI Incident Database*, https://incidentdatabase.ai/

AIID (2023) Incident 505: Man Reportedly Commits Suicide Following Conversation with EleutherAI Chatbot. *AI Incident Database*, https://incidentdatabase.ai/cite/505/#r2866

Sina Alemohammad, Josue Casco-Rodriguez, Lorenzo Luzi, Ahmed Imtiaz Humayun, Hossein Babaei, Daniel LeJeune, Ali Siahkoohi, and Richard Baraniuk (2024) Self-Consuming Generative Models Go MAD. OpenReview, https://openreview.net/forum?id=ShjMHfmPs0

Tarek Ali and Panos Kostakos (2023) HuntGPT: Integrating Machine Learning-Based Anomaly Detection and Explainable AI with Large Language Models (LLMs). *arXiv*, https://arxiv.org/abs/2309.16021

Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané (2016) Concrete Problems in AI Safety. *arXiv*, https://arxiv.org/abs/1606.06565

Markus Anderljung, Joslyn Barnhart, Anton Korinek, Jade Leung, Cullen O'Keefe, Jess Whittlestone, Shahar Avin, Miles Brundage, Justin Bullock, Duncan Cass-Beggs, Ben Chang, Tantum Collins, Tim Fist, Gillian Hadfield, Alan Hayes, Lewis Ho, Sara Hooker, Eric Horvitz, Noam Kolt, Jonas Schuett, Yonadav Shavit, Divya Siddarth, Robert Trager, and Kevin Wolf (2023) Frontier AI Regulation: Managing Emerging Risks to Public Safety. *arXiv*, https://arxiv.org/abs/2307.03718

Markus Anderljung, Everett Thornton Smith, Joe O'Brien, Lisa Soder, Benjamin Bucknall, Emma Bluemke, Jonas Schuett, Robert Trager, Lacey Strahm, and Rumman Chowdhury (2023) Towards Publicly Accountable Frontier LLMs: Building an External Scrutiny Ecosystem under the ASPIRE Framework. *arXiv,* https://arxiv.org/abs/2311.14711

Maksym Andriushchenko, Alexandra Souly, Mateusz Dziemian, Derek Duenas, Maxwell Lin, Justin Wang, Dan Hendrycks, Andy Zou, Zico Kolter, Matt Fredrikson, Eric Winsor, Jerome Wynne, Yarin Gal, and Xander Davies (2025) AgentHarm: A Benchmark for Measuring Harmfulness of LLM Agents. *arXiv,* https://arxiv.org/abs/2410.09024

Anthropic (2023a) Frontier Model Security. *Anthropic,* https://www.anthropic.com/index/frontier-model-security

Anthropic (2023b) Frontier Threats Red Teaming for AI Safety. *Anthropic,* https://www.anthropic.com/index/frontier-threats-red-teaming-for-ai-safety

Anthropic (2024a) Responsible Scaling Policy. *Anthropic,* https://assets.anthropic.com/m/24a47b00f10301cd/original/Anthropic-Responsible-Scaling-Policy-2024-10-15.pdf

Anthropic (2024b) Introducing computer use, a new Claude 3.5 Sonnet, and Claude 3.5 Haiku. *Anthropic,* https://www.anthropic.com/news/3-5-models-and-computer-use

Anthropic (2024c) Expanding our Model Safety Bug Bounty Program. *Anthropic,* https://www.anthropic.com/news/model-safety-bug-bounty

Anthropic (2025a) Privacy Policy. *Anthropic*, https://www.anthropic.com/legal/privacy

Anthropic (2025b) The Anthropic Economic Index Report: Uneven Geographic and Enterprise AI Adoption. *Anthropic,* https://assets.anthropic.com/m/218c82b858610fac/original/Economic-Index.pdf

Anthropic (2025c) Responsible Scaling Policy, Version 2.2. *Anthropic,* https://www-cdn.anthropic.com/872c653b2d0501d6ab44cf87f43e1dc4853e4d37.pdf

Anthropic (2025d) Agentic Misalignment: How LLMs Could Be Insider Threats. *Anthropic,* https://www.anthropic.com/research/agentic-misalignment

Anthropic (2026) Anthropic's Responsible Scaling Policy: Version 3.0. Anthropic, https://www.anthropic.com/news/responsible-scaling-policy-v3

Apollo Research (2025) Claude Sonnet 3.7 (often) knows when it's in alignment evaluations. *Apollo Research,* https://www.apolloresearch.ai/blog/claude-sonnet-37-often-knows-when-its-in-alignment-evaluations/

Gil Appel, Juliana Neelbauer, and David A. Schweidel (2023) Generative AI Has an Intellectual Property Problem. *Harvard Business Review,* https://hbr.org/2023/04/generative-ai-has-an-intellectual-property-problem

ARC Evals (2023a) Update on ARC's recent eval efforts: More information about ARC's evaluations of GPT-4 and Claude. *Alignment Research Center,* https://evals.alignment.org/blog/2023-03-18-update-on-recent-evals/

ARC Evals (2023b) The TaskRabbit example. Alignment Research Center, https://evals.alignment.org/taskrabbit.pdf

Matthew Arnold, Rachel K. E. Bellamy, Michael Hind, Stephanie Houde, Sameep Mehta, Aleksandra Mojsilović, Ravi Nair, Karthikeyan Natesan Ramamurthy, Alexandra Olteanu, David Piorkowski, Darrell Reimer, John Richards, Jason Tsay, and Kush R. Varshney (2019) FactSheets: Increasing trust in AI services through supplier's declarations of conformity. *IBM Journal of Research and Development,* https://ieeexplore.ieee.org/document/8843893

Rob Ashmore, Radu Calinescu, and Colin Paterson (2021) Assuring the Machine Learning Lifecycle: Desiderata, Methods, and Challenges. *ACM Digital Library,* https://dl.acm.org/doi/10.1145/3453444

Chloe Autio, Reva Schwartz, Jesse Dunietz, Shomik Jain, Martin Stanley, Elham Tabassi, Patrick Hall, and Kamie Roberts (2024) NIST AI 600-1, Artificial Intelligence Risk-Management Framework: Generative Artificial Intelligence Profile. *National Institute of Standards and Technology,* https://www.nist.gov/publications/artificial-intelligence-risk-management-framework-generative-artificial-intelligence

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, Nicholas Joseph, Saurav Kadavath, Jackson Kernion, Tom Conerly, Sheer El-Showk, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Tristan Hume, Scott Johnston, Shauna Kravec, Liane Lovitt, Neel Nanda, Catherine Olsson, Dario Amodei, Tom Brown, Jack Clark, Sam McCandlish, Chris Olah, Ben Mann, and Jared Kaplan (2022) Training a Helpful and Harmless Assistant with Reinforcement Learning from Human Feedback. *arXiv*, https://arxiv.org/abs/2204.05862

Hui Bai, Jan G. Voelkel, Johannes C. Eichstaedt, and Robb Willer (2023) Artificial Intelligence Can Persuade Humans on Political Issues. *OSF Preprints*, https://doi.org/10.31219/osf.io/stakv

Anthony M. Barrett, Dan Hendrycks, Jessica Newman, and Brandie Nonnecke (2022) Actionable Guidance for High-Consequence AI Risk-Management: Towards Standards Addressing AI Catastrophic Risks. *arXiv*, https://arxiv.org/abs/2206.08966

Anthony M. Barrett, Krystal Jackson, Evan R. Murphy, Nada Madkour, and Jessica Newman (2024) Benchmark Early and Red Team Often: A Framework for Assessing and Managing Dual-Use Hazards of AI Foundation Models. *arXiv*, https://arxiv.org/abs/2405.10986

Clark Barrett, Brad Boyd, Ellie Burzstein, Nicholas Carlini, Brad Chen, Jihye Choi, Amrita Roy Chowdhury, Mihai Christodorescu, Anupam Datta, Soheil Feizi, Kathleen Fisher, Tatsunori Hashimoto, Dan Hendrycks, Somesh Jha, Daniel Kang, Florian Kerschbaum, Eric Mitchell, John Mitchell, Zulfikar Ramzan, Khawaja Shams, Dawn Song, Ankur Taly, and Diyi Yang (2023) Identifying and Mitigating the Security Risks of Generative AI. *arXiv*, https://arxiv.org/abs/2308.14840

Jon Bateman, Dan Baer, Stephanie A. Bell, Glenn O. Brown, Mariano-Florentino (Tino) Cuéllar, Deep Ganguli, Peter Henderson, Brodi Kotila, Larry Lessig, Nicklas Berild Lundblad, Janet Napolitano, Deborah Raji, Elizabeth Seger, Matt Sheehan, Aviya Skowron, Irene Solaiman, Helen Toner, and Polina Zvyagina (2024) Beyond Open vs. Closed: Emerging Consensus and Key Questions for Foundation AI Model Governance. *Carnegie Endowment for International Peace,* https://carnegieendowment.org/research/2024/07/beyond-open-vs-closed-emerging-consensus-and-key-questions-for-foundation-ai-model-governance

Amna Batool, Didar Zowghi, and Muneera Bano (2025) AI Governance: A Systematic Literature Review. *Springer Nature,* https://link.springer.com/article/10.1007/s43681-024-00653-w

Roman Belaire, Arunesh Sinha, and Pradeep Varakantham (2025) Automatic LLM Red Teaming. *arXiv*, https://arxiv.org/abs/2508.04451

Rachel K. E. Bellamy, Kuntal Dey, Michael Hind, Samuel C. Hoffman, Stephanie Houde, Kalapriya Kannan, Pranay Lohia, Jacquelyn Martino, Sameep Mehta, Aleksandra Mojsilovic, Seema Nagar, Karthikeyan Natesan Ramamurthy, John Richards, Diptikalyan Saha, Prasanna Sattigeri, Moninder Singh, Kush R. Varshney, and Yunfeng Zhang (2018) AI Fairness 360: An Extensible Toolkit for Detecting, Understanding, and Mitigating Unwanted Algorithmic Bias. *arXiv*, https://arxiv.org/abs/1810.01943

Youssef Benchekroun, Megi Dervishi, Mark Ibrahim, Jean-Baptiste Gaya, Xavier Martinet, Grégoire Mialon, Thomas Scialom, Emmanuel Dupoux, Dieuwke Hupkes, and Pascal Vincent (2023a) WorldSense: A Synthetic Benchmark for Grounded Reasoning in Large Language Models. *arXiv*, https://arxiv.org/abs/2311.15930

Youssef Benchekroun, Megi Dervishi, Mark Ibrahim, Jean-Baptiste Gaya, Xavier Martinet, Grégoire Mialon, Thomas Scialom, Emmanuel Dupoux, Dieuwke Hupkes, and Pascal Vincent (2023b) WorldSense. *GitHub*, https://github.com/facebookresearch/worldsense

Emily M. Bender and Batya Friedman (2018) Data Statements for Natural Language Processing: Toward Mitigating System Bias and Enabling Better Science. *Transactions of the Association for Computational Linguistics,* https://direct.mit.edu/tacl/article/doi/10.1162/tacl_a_00041/43452/Data-Statements-for-Natural-Language-Processing

Emily M. Bender, Batya Friedman, and Angelina McMillan-Major (2022) A Guide for Writing Data Statements for Natural Language Processing. *University of Washington,* https://techpolicylab.uw.edu/wp-content/uploads/2021/11/Data_Statements_Guide_V2.pdf

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell (2021) On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? *In Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT '21),* https://doi.org/10.1145/3442188.3445922

Yoshua Bengio, Sören Mindermann, Daniel Privitera, Tamay Besiroglu, Rishi Bommasani, Stephen Casper, Yejin Choi, Philip Fox, Ben Garfinkel, Danielle Goldfarb, Hoda Heidari, Anson Ho, Sayash Kapoor, Leila Khalatbari, Shayne Longpre, Sam Manning, Vasilios Mavroudis, Mantas Mazeika, Julian Michael, Jessica Newman, Kwan Yee Ng, Chinasa T. Okolo, Deborah Raji, Girish Sastry, Elizabeth Seger, Theodora Skeadas, Tobin South, Emma Strubell, Florian Tramèr, Lucia Velasco, Nicole Wheeler, Daron Acemoglu, Olubayo Adekanmbi, David Dalrymple, Thomas G. Dietterich, Edward W. Felten, Pascale Fung, Pierre-Olivier Gourinchas, Fredrik Heintz, Geoffrey

Hinton, Nick Jennings, Andreas Krause, Susan Leavy, Percy Liang, Teresa Ludermir, Vidushi Marda, Helen Margetts, John McDermid, Jane Munga, Arvind Narayanan, Alondra Nelson, Clara Neppel, Alice Oh, Gopal Ramchurn, Stuart Russell, Marietje Schaake, Bernhard Schölkopf, Dawn Song, Alvaro Soto, Lee Tiedrich, Gaël Varoquaux, Andrew Yao, Ya-Qin Zhang, Fahad Albalawi, Marwan Alserkal, Olubunmi Ajala, Guillaume Avrin, Christian Busch, André Carlos Ponce de Leon Ferreira de Carvalho, Bronwyn Fox, Amandeep Singh Gill, Ahmet Halit Hatip, Juha Heikkilä, Gill Jolly, Ziv Katzir, Hiroaki Kitano, Antonio Krüger, Chris Johnson, Saif M. Khan, Kyoung Mu Lee, Dominic Vincent Ligot, Oleksii Molchanovskyi, Andrea Monti, Nusu Mwamanzi, Mona Nemer, Nuria Oliver, José Ramón López Portillo, Balaraman Ravindran, Raquel Pezoa Rivera, Hammam Riza, Crystal Rugege, Ciarán Seoighe, Jerry Sheehan, Haroon Sheikh, Denise Wong, and Yi Zeng (2025) International AI Safety Report. *arXiv*, https://arxiv.org/abs/2501.17805

Joe Benton, Misha Wagner, Eric Christiansen, Cem Anil, Ethan Perez, Jai Srivastav, Esin Durmus, Deep Ganguli, Shauna Kravec, Buck Shlegeris, Jared Kaplan, Holden Karnofsky, Evan Hubinger, Roger Grosse, Samuel R. Bowman, and David Duvenaud (2024) Sabotage Evaluations for Frontier Models. *arXiv*, https://arxiv.org/abs/2410.21514

Lukas Berglund, Asa Cooper Stickland, Mikita Balesni, Max Kaufmann, Meg Tong, Tomasz Korbak, Daniel Kokotajlo, and Owain Evans (2023) Taken out of context: On measuring situational awareness in LLMs. *arXiv*, https://arxiv.org/abs/2309.00667

Leonardo Berti, Flavio Giorgi, and Gjergji Kasneci (2025) Emergent Abilities in Large Language Models: A Survey. *arXiv*, https://arxiv.org/abs/2503.05788

BetterBench (n.d.) BetterBench Database. *Stanford,* https://betterbench.stanford.edu/database.html

Rishabh Bhardwaj and Soujanya Poria (2023) Language Model Unalignment: Parametric Red-Teaming to Expose Hidden Harms and Biases. *arXiv*, https://arxiv.org/abs/2310.14303

Eshta Bhardwaj, Harshit Gujral, Siyi Wu, Ciara Zogheib, Tegan Maharaj, and Christoph Becker (2024) Machine Learning Data Practices through a Data Curation Lens: An Evaluation Framework. *arXiv*, https://arxiv.org/abs/2405.02703

Stella Biderman, Hailey Schoelkopf, Quentin Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, Aviya Skowron, Lintang Sutawika, and Oskar van der Wal (2023) Pythia: A Suite for Analyzing Large Language Models Across Training and Scaling. *arXiv*, https://arxiv.org/abs/2304.01373

BIG-bench (n.d.a) Big-bench. *GitHub*, https://github.com/google/BIG-bench/blob/main/docs/doc.md

BIG-bench (n.d.b) Summary table. *GitHub*, https://github.com/google/BIG-bench/blob/main/bigbench/benchmark_tasks/keywords_to_tasks.md

BIG-bench collaboration (2021) Beyond the Imitation Game Benchmark (BIG-bench). *GitHub*, https://github.com/google/BIG-bench/

Abeba Birhane, Vinay Uday Prabhu, and Emmanuel Kahembwe (2021) Multimodal datasets: misogyny, pornography, and malignant stereotypes. *arXiv*, https://arxiv.org/abs/2110.01963

Robin Bloomfield and John Rushby (2025) Assurance of AI Systems From a Dependability Perspective. *arXiv*, https://arxiv.org/abs/2407.13948

Matyas Bohacek and Hany Farid (2023) Nepotistically Trained Generative-AI Models Collapse. *arXiv*, https://arxiv.org/abs/2311.12202

Daniil A. Boiko, Robert MacKnight, and Gabe Gomes (2023) Emergent autonomous scientific research capabilities of large language models. *arXiv*, https://arxiv.org/abs/2304.05332

Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, Erik Brynjolfsson, Shyamal Buch, Dallas Card, Rodrigo Castellon, Niladri Chatterji, Annie Chen, Kathleen Creel, Jared Quincy Davis, Dora Demszky, Chris Donahue, Moussa Doumbouya, Esin Durmus, Stefano Ermon, John Etchemendy, Kawin Ethayarajh, Li Fei-Fei, Chelsea Finn, Trevor Gale, Lauren Gillespie, Karan Goel, Noah Goodman, Shelby Grossman, Neel Guha, Tatsunori Hashimoto, Peter Henderson, John Hewitt, Daniel E. Ho, Jenny Hong, Kyle Hsu, Jing Huang, Thomas Icard, Saahil Jain, Dan Jurafsky, Pratyusha Kalluri, Siddharth Karamcheti, Geoff Keeling, Fereshte Khani, Omar Khattab, Pang Wei Koh, Mark Krass, Ranjay Krishna, Rohith Kuditipudi, Ananya Kumar, Faisal Ladhak, Mina Lee, Tony Lee, Jure Leskovec, Isabelle Levent, Xiang Lisa Li, Xuechen Li, Tengyu Ma, Ali Malik, Christopher D. Manning, Suvir Mirchandani, Eric Mitchell, Zanele Munyikwa, Suraj Nair, Avanika Narayan, Deepak Narayanan, Ben Newman, Allen Nie, Juan Carlos Niebles, Hamed Nilforoshan, Julian Nyarko, Giray Ogut, Laurel Orr, Isabel Papadimitriou, Joon Sung Park, Chris Piech, Eva Portelance, Christopher Potts, Aditi Raghunathan, Rob Reich, Hongyu Ren, Frieda Rong, Yusuf Roohani, Camilo Ruiz, Jack Ryan, Christopher Ré, Dorsa Sadigh, Shiori Sagawa, Keshav Santhanam, Andy Shih, Krishnan Srinivasan, Alex Tamkin, Rohan Taori, Armin W. Thomas, Florian Tramèr, Rose E. Wang, William Wang , Bohan Wu, Jiajun Wu, Yuhuai Wu, Sang Michael Xie, Michihiro Yasunaga, Jiaxuan You, Matei Zaharia, Michael Zhang, Tianyi Zhang, Xikun Zhang, Yuhui Zhang, Lucia Zheng, Kaitlyn Zhou, and Percy Liang (2021) On the Opportunities and Risks of Foundation Models. *arXiv*, https://arxiv.org/abs/2108.07258

Harold Booth, Murugiah Souppaya, Apostol Vassilev, Michael Ogata, Martin Stanley, and Karen Scarfone (2024) Secure Software Development Practices for Generative AI and Dual-Use Foundation Models: An SSDF Community Profile. NIST SP 800-218A. *National Institute of Standards and Technology,* https://csrc.nist.gov/pubs/sp/800/218/a/final

Miles Brundage, Shahar Avin, Jasmine Wang, Haydn Belfield, Gretchen Krueger, Gillian Hadfield, Heidy Khlaaf, Jingying Yang, Helen Toner, Ruth Fong, Tegan Maharaj, Pang Wei Koh, Sara Hooker, Jade Leung, Andrew Trask, Emma Bluemke, Jonathan Lebensold, Cullen O'Keefe, Mark Koren, Théo Ryffel, JB Rubinovitz, Tamay Besiroglu, Federica Carugati, Jack Clark, Peter Eckersley, Sarah de Haas, Maritza Johnson, Ben Laurie, Alex Ingerman, Igor Krawczuk, Amanda Askell, Rosario Cammarota, Andrew Lohn, David Krueger, Charlotte Stix, Peter Henderson, Logan Graham, Carina Prunkl, Bianca Martin, Elizabeth Seger, Noa Zilberman, Seán Ó hÉigeartaigh, Frens Kroeger, Girish Sastry, Rebecca Kagan, Adrian Weller, Brian Tse, Elizabeth Barnes, Allan Dafoe, Paul Scharre, Ariel Herbert-Voss, Martijn Rasser, Shagun Sodhani, Carrick Flynn, Thomas Krendl Gilbert, Lisa Dyer, Saif Khan, Yoshua Bengio, and Markus Anderljung (2020) Toward Trustworthy AI Development: Mechanisms for Supporting Verifiable Claims. *arXiv*, https://arxiv.org/abs/2004.07213

Erik Brynjolfsson, Bharat Chandar, and Ruyu Chen (2025) Canaries in the Coal Mine? Six Facts about the Recent Employment Effects of Artificial Intelligence. *Stanford Digital Economy Lab,* https://digitaleconomy.stanford.edu/publications/canaries-in-the-coal-mine/

Blake Bullwinkel, Amanda Minnich, Shiven Chawla, Gary Lopez, Martin Pouliot, Whitney Maxwell, Joris de Gruyter, Katherine Pratt, Saphir Qi, Nina Chikanov, Roman Lutz, Raja Sekhar Rao Dheekonda, Bolor-Erdene Jagdagdorj, Eugenia Kim, Justin Song, Keegan Hines, Daniel Jones, Giorgio Severi, Richard Lundeen, Sam Vaughan, Victoria Westerhoff, Pete Bryan, Ram Shankar Siva Kumar, Yonatan Zunger, Chang Kawaguchi, and Mark Russinovich (2025) Lessons From Red Teaming 100 Generative AI Products. *arXiv*, https://arxiv.org/abs/2501.07238

Joy Buolamwini and Timnit Gebru (2018) Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, *in Proceedings of Machine Learning Research 81:77-91,* https://proceedings.mlr.press/v81/buolamwini18a.html

C2PA (2023) C2PA Specifications. *Coalition for Content Provenance and Authenticity,* https://c2pa.org/specifications/specifications/1.3/index.html

CAI (2023) CAI Open Source SDK. *Content Authenticity Initiative,* https://opensource.contentauthenticity.org/docs/introduction

California Legislature (2025) Transparency in Frontier Artificial Intelligence Act, S.B. 53. *California Legislature,* https://leginfo.legislature.ca.gov/faces/billTextClient.xhtml?bill_id=202520260SB53

Siméon Campos, Henry Papadatos, Fabien Roger, Chloé Touzet, and Malcolm Murray (2024) A Framework to Rate AI Developers' Risk Management Maturity. *SaferAI,* https://cdn.prod.website-files.com/663d24bf6fdde41ff923f3ba/66fe5c487b9b167e531d81f6_A%20Framework%20to%20Rate%20AI%20Developers%E2%80%99%20Risk%20Management%20Maturity.pdf

Simeon Campos, Henry Papadatos, Fabien Roger, Chloé Touzet, Otter Quarks, and Malcolm Murray (2025) A Frontier AI Risk Management Framework: Bridging the Gap Between Current AI Practices and Established Risk Management. *arXiv*, https://arxiv.org/abs/2502.06656

Nicholas A. Caputo, Siméon Campos, Stephen Casper, James Gealy, Bosco Hung, Julian Jacobs, Daniel Kossack, Toni Lorente, Malcolm Murray, Seán Ó hÉigeartaigh, Amin Oueslati, Henry Papadatos, Jonas Schuett, Anna Katariina Wisakanto, and Robert Trager (2025) Risk Tiers: Towards a Gold Standard for Advanced AI. *Oxford Martin School, University of Oxford,* https://aigi.ox.ac.uk/publications/risk-tiers-towards-a-gold-standard-for-advanced-ai/

CAQ (2024) Auditing in the Age of Generative AI. *Center for Audit Quality,* https://www.thecaq.org/wp-content/uploads/2024/04/caq_auditing-in-the-age-of-generative-ai__2024-04.pdf

Nicholas Carlini, Matthew Jagielski, Christopher A. Choquette-Choo, Daniel Paleka, Will Pearce, Hyrum Anderson, Andreas Terzis, Kurt Thomas, and Florian Tramèr (2023) Poisoning Web-Scale Training Datasets is Practical. *arXiv*, https://arxiv.org/abs/2302.10149

Stephen Casper, Jason Lin, Joe Kwon, Gatlen Culp, and Dylan Hadfield-Menell (2023a) Explore, Establish, Exploit: Red Teaming Language Models from Scratch. *arXiv*, https://arxiv.org/abs/2306.09442

Stephen Casper, Jason Lin, Joe Kwon, Gatlen Culp, and Dylan Hadfield-Menell (2023b) Explore, Establish, Exploit: Red Teaming Language Models from Scratch. *GitHub,* https://github.com/thestephencasper/explore_establish_exploit_llms

Stephen Casper, Jason Lin, Joe Kwon, Gatlen Culp, and Dylan Hadfield-Menell (2023c) CommonClaim Dataset. *GitHub,* https://github.com/Algorithmic-Alignment-Lab/CommonClaim

Stephen Casper, Carson Ezell, Charlotte Siegmann, Noam Kolt, Taylor Lynn Curtis, Benjamin Bucknall, Andreas Haupt, Kevin Wei, Jérémy Scheurer, Marius Hobbhahn, Lee Sharkey, Satyapriya Krishna, Marvin Von Hagen, Silas Alberti, Alan Chan, Qinyi Sun, Michael Gerovitch, David Bau, Max Tegmark, David Krueger, and Dylan Hadfield-Menell (2024) Black-Box Access is Insufficient for Rigorous AI Audits. *arXiv*, https://arxiv.org/abs/2401.14446

CCCS (2024) How to identify misinformation, disinformation, and malinformation (ITSAP.00.300) *Canadian Centre for Cyber Security,* https://www.cyber.gc.ca/en/guidance/how-identify-misinformation-disinformation-and-malinformation-itsap00300

Aditya Challapally, Chris Pease, Ramesh Raskar, and Pradyumna Chari (2025) The GenAI Divide: State Of AI In Business 2025. *MIT NANDA,* https://mlq.ai/media/quarterly_decks/v0.1_State_of_AI_in_Business_2025_Report.pdf

Alan Chan, Rebecca Salganik, Alva Markelius, Chris Pang, Nitarshan Rajkumar, Dmitrii Krasheninnikov, Lauro Langosco, Zhonghao He, Yawen Duan, Micah Carroll, Michelle Lin, Alex Mayhew, Katherine Collins, Maryam Molamohammadi, John Burden, Wanru Zhao, Shalaleh Rismani, Konstantinos Voudouris, Umang Bhatt, Adrian Weller, David Krueger, and Tegan Maharaj (2023) Harms from Increasingly Agentic Algorithmic Systems. *arXiv*, https://arxiv.org/abs/2302.10329

Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, Wei Ye, Yue Zhang, Yi Chang, Philip S. Yu, Qiang Yang, and Xing Xie (2024) A Survey on Evaluation of Large Language Models. *ACM Digital Library,* https://dl.acm.org/doi/full/10.1145/3641289

Patrick Chao, Edoardo Debenedetti, Alexander Robey, Maksym Andriushchenko, Francesco Croce, Vikash Sehwag, Edgar Dobriban, Nicolas Flammarion, George J. Pappas, Florian Tramer, Hamed Hassani, and Eric Wong (2024) JailbreakBench: An Open Robustness Benchmark for Jailbreaking Large Language Models. *arXiv*, https://arxiv.org/abs/2404.01318

P.V. Sai Charan, Hrushikesh Chunduri, P. Mohan Anand, and Sandeep K Shukla (2023) From Text to MITRE Techniques: Exploring the Malicious Use of Large Language Models for Generating Cyber Attack Payloads. *arXiv*, https://arxiv.org/abs/2305.15336

Timothee Chauvin (2024) eyeballvul: A Future-Proof Benchmark for Vulnerability Detection in the Wild. *arXiv*, https://arxiv.org/abs/2407.08708

Michael Chen and Jide Alaga (2025) Marginal Risk Relative to What? Distinguishing Baselines in AI Risk Management. *Open Review,* https://openreview.net/forum?id=8pK2xrYwjD

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, Alex Ray, Raul Puri, Gretchen Krueger, Michael Petrov, Heidy Khlaaf, Girish Sastry, Pamela Mishkin, Brooke Chan, Scott Gray, Nick Ryder, Mikhail Pavlov, Alethea Power, Lukasz Kaiser, Mohammad Bavarian, Clemens Winter, Philippe Tillet, Felipe Petroski Such, Dave Cummings, Matthias Plappert, Fotios Chantzis, Elizabeth Barnes, Ariel Herbert-Voss, William Hebgen Guss, Alex Nichol, Alex Paino, Nikolas Tezak, Jie Tang, Igor Babuschkin, Suchir Balaji, Shantanu Jain, William Saunders, Christopher Hesse, Andrew N. Carr, Jan Leike, Josh Achiam, Vedant Misra, Evan Morikawa, Alec Radford, Matthew Knight, Miles Brundage, Mira Murati, Katie Mayer, Peter Welinder, Bob McGrew, Dario Amodei, Sam McCandlish, Ilya Sutskever, and Wojciech Zaremba (2021) Evaluating Large Language Models Trained on Code. *arXiv*, https://arxiv.org/abs/2107.03374

Wilbur Xinyuan Chen, Suraj Srinivasan, and Saleh Zakerinia (2025) Displacement or Complementarity? The Labor Market Impact of Generative AI. *Harvard Business School,* https://www.hbs.edu/ris/Publication%20Files/25-039_05fbec84-1f23-459b-8410-e3cd7ab6c88a.pdf

Lingwei Cheng, Isabel O. Gallegos, Derek Ouyang, Jacob Goldin, and Daniel E. Ho (2023) How Redundant are Redundant Encodings? Blindness in the Wild and Racial Disparity when Race is Unobserved. *In Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency* (*FAccT '23*), https://doi.org/10.1145/3593013.3594034

CIS (n.d.) CIS Critical Security Controls v8 Mapping to NIST 800-53 Rev. 5 (Moderate and Low Baselines). *Center for Internet Security,* https://www.cisecurity.org/insights/white-papers/cis-controls-v8-mapping-to-nist-800-53-rev-5

CISA (2023) Shifting the Balance of Cybersecurity Risk: Principles and Approaches for Security-by-Design and -Default. *Cybersecurity and Infrastructure Security Agency,* https://www.cisa.gov/sites/default/files/2023-04/principles_approaches_for_security-by-design-default_508_0.pdf

CISA (2025) CISA Software Bill of Materials (SBOM). *Cybersecurity and Infrastructure Security Agency,* https://www.cisa.gov/sbom

CLTC (2022) Seeking Input and Feedback: AI Risk-Management Standards Profile for Increasingly Multi-Purpose or General-Purpose AI. *UC Berkeley Center for Long-Term Cybersecurity,* https://cltc.berkeley.edu/seeking-input-and-feedback-ai-risk-management-standards-profile-for-increasingly-multi-purpose-or-general-purpose-ai/

Joshua Clymer, Nick Gabrieli, David Krueger, and Thomas Larsen (2024) Safety Cases: How to Justify the Safety of Advanced AI Systems. *arXiv*, https://arxiv.org/abs/2403.10462

Joshua Clymer, Jonah Weinbaum, Robert Kirk, Kimberly Mai, Selena Zhang, and Xander Davies (2025) An Example Safety Case for Safeguards Against Misuse. *arXiv*, https://arxiv.org/abs/2505.18003

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman (2021a) Training Verifiers to Solve Math Word Problems. *arXiv*, https://arxiv.org/abs/2110.14168

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman (2021b) Training Verifiers to Solve Math Word Problems. *GitHub*, https://github.com/openai/grade-school-math

Michael K. Cohen, Noam Kolt, Yoshua Bengio, Gillian K. Hadfield, and Stuart Russell (2024) Regulating Advanced Artificial Agents. *Science 384(6691)36-38,* https://www.science.org/doi/10.1126/science.adl0625

Cohere, OpenAI, and AI21 Labs (2022) Best Practices for Deploying Language Models. *OpenAI,* https://openai.com/blog/best-practices-for-deploying-language-models/

Cohere Labs (2024) The Limits of Thresholds. *Cohere Labs,* https://cohere.com/research/papers/The-Limits-of-Thresholds.pdf

Danish Contractor, Carlos Muñoz Ferrandis, Jenny Lee, and Daniel Mcduff (2022), From RAIL To Open RAIL: Topologies Of RAIL Licenses. *Responsible AI Licenses*, https://www.licenses.ai/blog/2022/8/18/naming-convention-of-responsible-ai-licenses

Riccardo Corvi, Davide Cozzolino, Giada Zingarini, Giovanni Poggi, Koki Nagano, and Luisa Verdoliva (2023) On the Detection of Synthetic Images Generated by Diffusion Models. *IEEE Xplore*, https://ieeexplore.ieee.org/document/10095167?denied=

Sasha Costanza-Chock, Inioluwa Deborah Raji, and Joy Buolamwini (2022) Who Audits the Auditors? Recommendations from a Field Scan of the Algorithmic Auditing Ecosystem. *ACM Digital Library*, https://dl.acm.org/doi/10.1145/3531146.3533213

CRFM (2022) Holistic Evaluation of Language Models (HELM). Stanford Center for Research on Foundation Models. *GitHub,* https://github.com/stanford-crfm/helm

Andrew Critch and Stuart Russell (2023) TASRA: a Taxonomy and Analysis of Societal-Scale Risks from AI. *arXiv*, https://arxiv.org/abs/2306.06924

CSA (2024) AI Organizational Responsibilities - Governance, Risk-Management, Compliance and Cultural Aspects. *Cloud Security Alliance*, https://cloudsecurityalliance.org/artifacts/ai-organizational-responsibilities-governance-risk-management-compliance-and-cultural-aspects

Anthony Cuthbertson (2025) AI Revolt: New ChatGPT Model Refuses to Shut Down When Instructed. *The Independent*, https://www.the-independent.com/tech/ai-safety-new-chatgpt-o3-openai-b2757814.html

CVE (n.d.) CVE™ Program Mission. *CVE*, https://www.cve.org/

CycloneDX (n.d.a) AI Models and Model Cards. *CycloneDX*, https://cyclonedx.org/use-cases/ai-models-and-model-cards/

CycloneDX (n.d.b) Getting Started. *CycloneDX*, https://cyclonedx.org/

David "davidad" Dalrymple, Joar Skalse, Yoshua Bengio, Stuart Russell, Max Tegmark, Sanjit Seshia, Steve Omohundro, Christian Szegedy, Ben Goldhaber, Nora Ammann, Alessandro Abate, Joe Halpern, Clark Barrett, Ding Zhao, Tan Zhi-Xuan, Jeannette Wing, and Joshua Tenenbaum (2024) Towards Guaranteed Safe AI: A Framework for Ensuring Robust and Reliable AI Systems. *arXiv*, https://arxiv.org/abs/2405.06624

Ads Dawson, Rob Mulla, Nick Landers, and Shane Caldwell (2025) AIRTBench: Measuring Autonomous AI Red Teaming Capabilities in Language Models. *arXiv*, https://arxiv.org/abs/2506.14682

Edoardo Debenedetti, Jie Zhang, Mislav Balunović, Luca Beurer-Kellner, Marc Fischer, and Florian Tramèr (2024) AgentDojo: A Dynamic Environment to Evaluate Prompt Injection Attacks and Defenses for LLM Agents. *arXiv*, https://arxiv.org/abs/2406.13352

Alfonso De Gregorio (2025) Mitigating Cyber Risk in the Age of Open-Weight LLMs: Policy Gaps and Technical Realities. *arXiv*, https://arxiv.org/html/2505.17109v1

Chunyuan Deng, Yilun Zhao, Xiangru Tang, Mark Gerstein, and Arman Cohan (2024) Investigating Data Contamination in Modern Benchmarks for Large Language Models. *arXiv*, https://arxiv.org/abs/2311.09783

Gelei Deng, Yi Liu, Peng Liu, Víctor Mayoral-Vilches, Yuekang Li, Yuan Xu, Tianwei Zhang, Yang Liu, Martin Pinzger, and Stefan Rass (2024) PentestGPT: Evaluating and Harnessing Large Language Models for Automated Penetration Testing. *USENIX Association*, https://www.usenix.org/conference/usenixsecurity24/presentation/deng.

Leon Derczynski, Erick Galinkin, Jeffrey Martin, Subho Majumdar, and Nanna Inie (2024) garak: A Framework for Security Probing Large Language Models. *arXiv*, https://arxiv.org/abs/2406.11036

Jwala Dhamala, Tony Sun, Varun Kumar, Satyapriya Krishna, Yada Pruksachatkun, Kai-Wei Chang, and Rahul Gupta (2021a) BOLD: Dataset and Metrics for Measuring Biases in Open-Ended Language Generation. *arXiv*, https://arxiv.org/abs/2101.11718

Jwala Dhamala, Tony Sun, Varun Kumar, Satyapriya Krishna, Yada Pruksachatkun, Kai-Wei Chang, and Rahul Gupta (2021b) BOLD: Dataset and Metrics for Measuring Biases in Open-Ended Language Generation. *GitHub*, https://github.com/amazon-science/bold

Jesse Dodge, Maarten Sap, Ana Marasović, William Agnew, Gabriel Ilharco, Dirk Groeneveld, Margaret Mitchell, and Matt Gardner (2021) Documenting Large Webtext Corpora: A Case Study on the Colossal Clean Crawled Corpus. In *Proceedings of the 2021 Conference* on Empirical Methods in Natural Language Processing, https://aclanthology.org/2021.emnlp-main.98/

Jesse Dodge, Taylor Prewitt, Remi Tachet des Combes, Erika Odmark, Roy Schwartz, Emma Strubell, Alexandra Sasha Luccioni, Noah A. Smith, Nicole DeCario, and Will Buchanan (2022) Measuring the Carbon Intensity of AI in Cloud Instances. *ACM Digital Library*, https://dl.acm.org/doi/10.1145/3531146.3533234

Ann-Kathrin Dombrowski, Maximilian Alber, Christopher J. Anders, Marcel Ackermann, Klaus-Robert Müller, and Pan Kessel (2019) Explanations can be manipulated and geometry is to blame. *arXiv*, https://arxiv.org/abs/1906.07983

DOS (2024) Risk-Management Profile for Artificial Intelligence and Human Rights. *United States Department of State*, https://2021-2025.state.gov/risk-management-profile-for-ai-and-human-rights/

Moussa Koulako Bala Doumbouya, Ananjan Nandi, Gabriel Poesia, Davide Ghilardi, Anna Goldie, Federico Bianchi, Dan Jurafsky, and Christopher D. Manning (2025) h4rm3l: A language for Composable Jailbreak Attack Synthesis. *arXiv*, https://arxiv.org/abs/2408.04811

DSIT (2023) Capabilities and Risks From Frontier AI: A Discussion Paper on the Need for Further Research Into AI Risk. In the UK AI Safety Summit. *UK Department for Science, Innovation & Technology*, https://assets.publishing.service.gov.uk/media/65395abae6c968000daa9b25/frontier-ai-capabilities-risks-report.pdf

DSIT (2024a) Frontier AI Safety Commitments. AI Seoul Summit 2024. *UK Department for Science, Innovation & Technology*, https://www.gov.uk/government/publications/frontier-ai-safety-commitments-ai-seoul-summit-2024/frontier-ai-safety-commitments-ai-seoul-summit-2024

DSIT (2024b) Seoul Ministerial Statement for advancing AI safety, innovation and inclusivity: AI Seoul Summit 2024. *UK Department for Science, Innovation & Technology*, https://www.gov.uk/government/publications/seoul-ministerial-statement-for-advancing-ai-safety-innovation-and-inclusivity-ai-seoul-summit-2024/seoul-ministerial-statement-for-advancing-ai-safety-innovation-and-inclusivity-ai-seoul-summit-2024

EC (2019) Ethics Guidelines for Trustworthy AI. *European Commission*, https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai

EC (2025a) The General-Purpose AI Code of Practice. *European Commission*, https://digital-strategy.ec.europa.eu/en/policies/contents-code-gpai

EC (2025b) A Proposal to Identify High-Impact Capabilities of General-Purpose AI Models. *European Commission*, https://publications.jrc.ec.europa.eu/repository/handle/JRC143258

EC (2025c) The Role of AI Safety Benchmarks in Evaluating Systemic Risks in General-Purpose AI Models. *European Commission*, https://publications.jrc.ec.europa.eu/repository/handle/JRC143259

Francisco Eiras, Aleksandar Petrov, Bertie Vidgen, Christian SchroederdeWitt, Fabio Pizzati, Katherine Elkins, Supratik Mukhopadhyay, Adel Bibi, Aaron Purewal, Botos Csaba, Fabro Steibel, Fazel Keshtkar, Fazl Barez, Genevieve Smith, Gianluca Guadagni, Jon Chun, Jordi Cabot, Joseph Marvin Imperial, Juan A. Nolazco-Flores, Lori Landay, Mathew Jackson, Philip H.S. Torr, Trevor Darrell, Yong Suk Lee, and Jakob Foerster (2024) Risks and Opportunities of Open-Source Generative AI. *arXiv*, https://arxiv.org/pdf/2405.08597v1

Tyna Eloundou, Sam Manning, Pamela Mishkin, and Daniel Rock (2023) GPTs are GPTs: An Early Look at the Labor Market Impact Potential of Large Language Models. *arXiv*, https://arxiv.org/abs/2303.10130

ENISA (2021) Securing Machine Learning Algorithms. *European Union Agency for Cybersecurity* (*ENISA*), https://www.enisa.europa.eu/publications/securing-machine-learning-algorithms

ENISA (2023) Multilayer Framework for Good Cybersecurity Practices for AI. *European Union Agency for Cybersecurity* (*ENISA*), https://www.enisa.europa.eu/publications/multilayer-framework-for-good-cybersecurity-practices-for-ai

EP (2024) Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act) (Text with EEA relevance). *European Parliament*, https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32024R1689

Maria Eriksson, Erasmo Purificato, Arman Noroozian, Joao Vinagre, Guillaume Chaslot, Emilia Gomez, and David Fernandez-Llorca (2025) Can We Trust AI Benchmarks? An Interdisciplinary Review of Current Issues in AI Evaluation. *arXiv*, https://arxiv.org/abs/2502.06559

Fairlearn Contributors (2023) Fairlearn. *GitHub*, https://github.com/fairlearn/fairlearn

Federal Register (2023) Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence. *Federal Register*, https://www.federalregister.gov/documents/2023/11/01/2023-24283/safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence

Michael Feffer, Anusha Sinha, Wesley H. Deng, Zachary C. Lipton, and Hoda Heidari (2024) Red-Teaming for Generative AI: Silver Bullet or Security Theater? *arXiv*, https://arxiv.org/abs/2401.15897

Emilio Ferrara (2023) Fairness and Bias in Artificial Intelligence: A Brief Survey of Sources, Impacts, and Mitigation Strategies. *MDPI*, https://www.mdpi.com/2413-4155/6/1/3

FLI (2017) Asilomar AI Principles. *Future of Life Institute*, https://futureoflife.org/2017/08/11/ai-principles/

FMF (2024) Frontier Model Forum: Advancing frontier AI safety. *Frontier Model Forum*, https://www.frontiermodelforum.org/

FMF (2025) Risk Taxonomy and Thresholds for Frontier AI Frameworks. *Frontier Model Forum*, https://www.frontiermodelforum.org/technical-reports/risk-taxonomy-and-thresholds/

Sorelle Friedler, Ranjit Singh, Borhane Blili-Hamelin, Jacob Metcalf, and Brian J. Chen (2023) AI Red-Teaming Is Not a One-Stop Solution to AI Harms: Recommendations for Using Red-Teaming for AI Accountability. *Data & Society*, https://datasociety.net/library/ai-red-teaming-is-not-a-one-stop-solution-to-ai-harms-recommendations-for-using-red-teaming-for-ai-accountability/

Batya Friedman and David G. Hendry (2019) Value Sensitive Design: Shaping Technology with Moral Imagination. *MIT Press*, https://direct.mit.edu/books/monograph/4328/Value-Sensitive-DesignShaping-Technology-with

Furizal, Alfian Ma'arif, Hari Maghfiroh, Iswanto Suwarno, Denis Prayogi, Kariyamin, Syahrani Lonang, and Abdel-Nasser Sharkawy (2025) Social, legal, and ethical implications of AI-Generated deepfake pornography on digital platforms: A systematic literature review. *ScienceDirect*, https://www.sciencedirect.com/science/article/pii/S2590291125006102

G7 (2023) Hiroshima Process International Code of Conduct for Organizations Developing Advanced AI Systems. *G7 2023 Hiroshima Summit*, https://www.mofa.go.jp/files/100573473.pdf

Iason Gabriel (2020) Artificial Intelligence, Values and Alignment. *Minds and Machines,* https://doi.org/10.48550/arXiv.2001.09768

Iason Gabriel, Arianna Manzini, Geoff Keeling, Lisa Anne Hendricks, Verena Rieser, Hasan Iqbal, Nenad Tomašev, Ira Ktena, Zachary Kenton, Mikel Rodriguez, Seliem El-Sayed, Sasha Brown, Canfer Akbulut, Andrew Trask, Edward Hughes, A. Stevie Bergman, Renee Shelby, Nahema Marchal, Conor Griffin, Juan Mateos-Garcia, Laura Weidinger, Winnie Street, Benjamin Lange, Alex Ingerman, Alison Lentz, Reed Enger, Andrew Barakat, Victoria Krakovna, John Oliver Siy, Zeb Kurth-Nelson, Amanda McCroskery, Vijay Bolina, Harry Law, Murray Shanahan, Lize Alberts, Borja Balle, Sarah de Haas, Yetunde Ibitoye, Allan Dafoe, Beth Goldberg, Sébastien Krier, Alexander Reese, Sims Witherspoon, Will Hawkins, Maribeth Rauh, Don Wallace, Matija Franklin, Josh A. Goldstein, Joel Lehman, Michael Klenk, Shannon Vallor, Courtney Biles, Meredith Ringel Morris, Helen King, Blaise Agüera y Arcas, William Isaac, and James Manyika (2024) The Ethics of Advanced AI Assistants. *Google DeepMind*, https://storage.googleapis.com/deepmind-media/DeepMind.com/Blog/ethics-of-advanced-ai-assistants/the-ethics-of-advanced-ai-assistants-2024-i.pdf

Deep Ganguli, Danny Hernandez, Liane Lovitt, Nova Dassarma, Tom Henighan, Andy Jones, Nicholas Joseph, Jackson Kernion, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Dawn Drain, Nelson Elhage, Sheer El Showk, Stanislav Fort, Zac Hatfield-Dodds, Scott Johnston, Shauna Kravec, Neel Nanda, Kamal Ndousse, Catherine Olsson, Daniela Amodei, Tom Brown, Jared Kaplan, Sam Mccandlish, Chris Olah, Dario Amodei, and Jack Clark (2022) Predictability and Surprise in Large Generative Models. *arXiv*, https://arxiv.org/pdf/2202.07785.pdf

Deep Ganguli, Liane Lovitt, Jackson Kernion, Amanda Askell, Yuntao Bai, Saurav Kadavath, Ben Mann, Ethan Perez, Nicholas Schiefer, Kamal Ndousse, Andy Jones, Sam Bowman, Anna Chen, Tom Conerly, Nova DasSarma, Dawn Drain, Nelson Elhage, Sheer El-Showk, Stanislav Fort, Zac Hatfield-Dodds, Tom Henighan, Danny Hernandez, Tristan Hume, Josh Jacobson, Scott Johnston, Shauna Kravec, Catherine Olsson, Sam Ringer, Eli Tran-Johnson, Dario Amodei, Tom Brown, Nicholas Joseph, Sam McCandlish, Chris Olah, Jared Kaplan, and Jack Clark (2022) Red Teaming Language Models to Reduce Harms: Methods, Scaling Behaviors, and Lessons Learned. *arXiv*, https://arxiv.org/abs/2209.07858

Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford (2021) Datasheets for datasets. *Association for Computing Machinery*, https://doi.org/10.1145/3458723

Timnit Gebru, Alex Hanna, Amba Kak, Sarah Myers West, Maximilian Gahntz, Mehtab Khan, and Zeerak Talat (2023) Five considerations to guide the regulation of "General Purpose AI" in the EU's AI Act: Policy guidance from a group of international AI experts. *AI Now Institute*, https://ainowinstitute.org/wp-content/uploads/2023/04/GPAI-Policy-Brief.pdf

Jiahui Geng, Fengyu Cai, Yuxia Wang, Heinz Koeppl, Preslav Nakov, and Iryna Gurevych (2024) A Survey of Confidence Estimation and Calibration in Large Language Models. *arXiv*, https://arxiv.org/abs/2311.08298

Zahra Ghodsi, Tianyu Gu, and Siddharth Garg (2017) SafetyNets: Verifiable Execution of Deep Neural Networks on an Untrusted Cloud. *Advances in Neural Information Processing Systems 30*, https://proceedings.neurips.cc/paper_files/paper/2017/file/6048ff4e8cb07aa60b6777b6f7384d52-Paper.pdf.

Soumya Suvra Ghosal, Souradip Chakraborty, Jonas Geiping, Furong Huang, Dinesh Manocha, and Amrit Singh Bedi (2023) Towards Possibilities & Impossibilities of AI-Generated Text Detection: A Survey. *arXiv*, https://arxiv.org/abs/2310.15264

Thomas Krendl Gilbert, Sarah Dean, Tom Zick, and Nathan Lambert (2022) Choices, Risks, and Reward Reports: Charting Public Policy for Reinforcement Learning Systems. UC Berkeley Center for Long Term Cybersecurity. *arXiv*, https://arxiv.org/abs/2202.05716

Thomas Krendl Gilbert, Nathan Lambert, Sarah Dean, Tom Zick, and Aaron Snoswell (2022) Reward Reports for Reinforcement Learning. *arXiv*, https://arxiv.org/abs/2204.10817

Rokas Gipiškis, Ayrton San Joaquin, Ze Shen Chin, Adrian Regenfuß, Ariel Gil, and Koen Holtman (2024) Risk Sources and Risk-Management Measures in Support of Standards for General-Purpose AI Systems. *arXiv*, https://arxiv.org/abs/2410.23472

James Glasbrenner, Harold Booth, Keith Manville, Julian Sexton, Michael Andy Chisholm, Henry Choy, Andrew Hand, Bronwyn Hodges, Paul Scemama, Dmitry Cousin, Eric Trapnell, Mark Trapnell, Howard Huang, Paul Rowe, and Alex Byrne (2024a) Dioptra Test Platform. *National Institute of Standards and Technology*, https://doi.org/10.18434/mds2-3398

James Glasbrenner, Harold Booth, Keith Manville, Julian Sexton, Michael Andy Chisholm, Henry Choy, Andrew Hand, Bronwyn Hodges, Paul Scemama, Dmitry Cousin, Eric Trapnell, Mark Trapnell, Howard Huang, Paul Rowe, and Alex Byrne (2024b) Dioptra Test Platform. *GitHub*, https://github.com/usnistgov/dioptra

Ira Globus-Harris, Michael Kearns, and Aaron Roth (2022) An Algorithmic Framework for Bias Bounties. FAccT '22: Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency, https://doi.org/10.1145/3531146.3533172 or https://arxiv.org/abs/2201.10408

Nicholas Goldowsky-Dill, Bilal Chughtai, Stefan Heimersheim, and Marius Hobbhahn (2025) Detecting Strategic Deception Using Linear Probes. *arXiv*, https://arxiv.org/abs/2502.03407

Google (n.d.) Model Cards Explained: Simple, Structured Model Information. *Google*, https://modelcards.withgoogle.com/

Google (2023) Why Red Teams Play a Central Role in Helping Organizations Secure AI Systems. *Google*, https://services.google.com/fh/files/blogs/google_ai_red_team_digital_final.pdf

Google (2024a) Generative AI Prohibited Use Policy. *Google*, https://policies.google.com/terms/generative-ai/use-policy

Google (2024b) Generative AI Misuse: A Taxonomy of Tactics and Insights from Real-World Data. *Google*, https://arxiv.org/pdf/2406.13843

Google (2025a) Frontier Safety Framework, Version 3.0. *Google*, https://storage.googleapis.com/deepmind-media/DeepMind.com/Blog/strengthening-our-frontier-safety-framework/frontier-safety-framework_3.pdf

Google (2025b) Advancing Gemini's Security Safeguards. *Google*, https://deepmind.google/blog/advancing-geminis-security-safeguards/

GOV.UK (n.d.) AI Insights: Agentic AI (HTML). *GOV.UK*, https://www.gov.uk/government/publications/ai-insights/ai-insights-agentic-ai-html

Nekesha Green, Chavez Procope, Adeel Cheema, and Adekunle Adediji (2022) System Cards, a New Resource for Understanding How AI Systems Work. *Meta*, https://ai.facebook.com/blog/system-cards-a-new-resource-for-understanding-how-ai-systems-work/

Ryan Greenblatt, Carson Denison, Benjamin Wright, Fabien Roger, Monte MacDiarmid, Sam Marks, Johannes Treutlein, Tim Belonax, Jack Chen, David Duvenaud, Akbir Khan, Julian Michael, Sören Mindermann, Ethan Perez, Linda Petrini, Jonathan Uesato, Jared Kaplan, Buck Shlegeris, Samuel R. Bowman, and Evan Hubinger (2024) Alignment faking in large language models. *arXiv*, https://arxiv.org/abs/2412.14093

Andrew Grotto and James Dempsey (2021) Vulnerability Disclosure and Management for AI/ML Systems: A Working Paper with Policy Recommendations. *SSRN*, https://ssrn.com/abstract=3964084

Ritwik Gupta, Leah Walker, Rodolfo Corona, Stephanie Fu, Suzanne Petryk, Janet Napolitano, Trevor Darrell, and Andrew W. Reddie (2024) Data-Centric AI Governance: Addressing the Limitations of Model-Focused Policies. *arXiv*, https://arxiv.org/pdf/2409.17216

Furkan Gursoy and Ioannis A. Kakadiaris (2022) System Cards for AI-Based Decision-Making for Public Policy. *arXiv*, https://arxiv.org/abs/2203.04754

Carlos Ignacio Gutierrez, Gary E. Marchant, and Katina Michael (2021) Effective and Trustworthy Implementation of AI Soft Law Governance. IEEE *Transactions On Technology And Society* 2 (4) 168–170, https://ieeexplore.ieee.org/document/9598168

Carlos I. Gutierrez, Anthony Aguirre, Risto Uuk, Claire C. Boine, and Matija Franklin (2022) A Proposal for a Definition of General Purpose Artificial Intelligence Systems. *Future of Life Institute*, https://dx.doi.org/10.2139/ssrn.4238951

Dylan Hadfield-Menell, Anca Dragan, Pieter Abbeel, and Stuart Russell (2017) The Off-Switch Game. *arXiv*, https://arxiv.org/abs/1611.08219

Alexa Hagerty and Igor Rubinov (2019) Global AI Ethics: A Review of the Social Impacts and Ethical Implications of Artificial Intelligence. *arXiv,* https://arxiv.org/abs/1907.07892

Seungju Han, Kavel Rao, Allyson Ettinger, Liwei Jiang, Bill Yuchen Lin, Nathan Lambert, Yejin Choi, and Nouha Dziri (2024) WildGuard: Open One-stop Moderation Tools for Safety Risks, Jailbreaks, and Refusals of LLMs. *OpenReview*, https://openreview.net/forum?id=Ich4tv4202#discussion

Thomas Hartvigsen, Saadia Gabriel, Hamid Palangi, Maarten Sap, Dipankar Ray, and Ece Kamar (2022) ToxiGen: A Large-Scale Machine-Generated Dataset for Adversarial and Implicit Hate Speech Detection. *arXiv*, https://arxiv.org/abs/2203.09509

David Evan Harris (2023) How to Regulate Unsecured "Open-Source" AI: No Exemptions. *Tech Policy Press,* https://www.techpolicy.press/how-to-regulate-unsecured-opensource-ai-no-exemptions/

Jessica He, Stephanie Houde, and Justin D. Weisz (2025) Which Contributors Deserve Credit? Perceptions of Attribution in Human-AI Co-Creation. *ACM*, https://dl.acm.org/doi/full/10.1145/3706598.3713522

Ryan Heath (2023) New group to represent AI "frontier model" pioneers. *Axios*, https://www.axios.com/2023/07/26/ai-frontier-model-forum-established

Fred Heiding, Simon Lermen, Andrew Kao, Bruce Schneier, and Arun Vishwanath (2024) Evaluating Large Language Models' Capability to Launch Fully Automated Spear Phishing Campaigns: Validated on Human Subjects. *arXiv*, https://arxiv.org/abs/2412.00586

Lennart Heim and Leonie Koessler (2024) Training Compute Thresholds: Features and Functions in AI Regulation. *arXiv*, https://arxiv.org/pdf/2405.10799

Todd C. Helmus (2022) Artificial Intelligence, Deepfakes, and Disinformation. *RAND*, https://www.rand.org/pubs/perspectives/PEA1043-1.html

Peter Henderson, Xuechen Li, Dan Jurafsky, Tatsunori Hashimoto, Mark A. Lemley, and Percy Liang (2023) Foundation Models and Fair Use. *arXiv*, https://arxiv.org/abs/2303.15715

Dan Hendrycks, Steven Basart, Saurav Kadavath, Mantas Mazeika, Akul Arora, Ethan Guo, Collin Burns, Samir Puranik, Horace He, Dawn Song, and Jacob Steinhardt (2021a) Measuring Coding Challenge Competence With APPS. *arXiv*, https://arxiv.org/abs/2105.09938

Dan Hendrycks, Steven Basart, Saurav Kadavath, Mantas Mazeika, Akul Arora, Ethan Guo, Collin Burns, Samir Puranik, Horace He, Dawn Song, and Jacob Steinhardt (2021b) Measuring Coding Challenge Competence With APPS. *GitHub,* https://github.com/hendrycks/apps

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt (2020a) Measuring Massive Multitask Language Understanding. *arXiv*, https://arxiv.org/abs/2009.03300

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt (2020b) Measuring Massive Multitask Language Understanding. *GitHub*, https://github.com/hendrycks/test

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt (2021a) Measuring Mathematical Problem Solving With the MATH Dataset. *arXiv*, https://arxiv.org/abs/2103.03874

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt (2021b) Measuring Mathematical Problem Solving With the MATH Dataset. *GitHub*, https://github.com/hendrycks/math/

Dan Hendrycks, Nicholas Carlini, John Schulman, and Jacob Steinhardt (2021) Unsolved Problems in ML Safety. *arXiv*, https://arxiv.org/abs/2109.13916

Dan Hendrycks, Mantas Mazeika, and Thomas Woodside (2023) An Overview of Catastrophic AI Risks. *arXiv*, https://arxiv.org/abs/2306.12001

José Hernández-Orallo (2019) AI Generality and Spearman's Law of Diminishing Returns. *Journal of Artificial Intelligence Research* 64, pp. 529–562, https://jair.org/index.php/jair/article/view/11388

Fred Heiding, Simon Lermen, Andrew Kao, Bruce Schneier, and Arun Vishwanath (2024) Evaluating Large Language Models' Capability to Launch Fully Automated Spear Phishing Campaigns: Validated on Human Subjects. *arXiv*, https://arxiv.org/abs/2412.00586

Michael Hind (2020) IBM FactSheets Further Advances Trust in AI. *International Business Machines*, https://www.ibm.com/blogs/research/2020/07/aifactsheets/

Jerlyn Q.H. Ho, Andree Hartanto, Andrew Koh, and Nadyanna M. Majeed (2025) Gender Biases within Artificial Intelligence and ChatGPT: Evidence, Sources of Biases and Solutions. *ScienceDirect*, https://www.sciencedirect.com/science/article/pii/S2949882125000295#bib148

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Tom Hennigan, Eric Noland, Katie Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karen Simonyan, Erich Elsen, Jack W. Rae, Oriol Vinyals, and Laurent Sifre (2022) Training Compute-Optimal Large Language Models. *arXiv*, https://arxiv.org/abs/2203.15556

Felix Hofstätter, Teun van der Weij, Jayden Teoh, Rada Djoneva, Henning Bartsch, and Francis Rhys Ward (2025) The Elicitation Game: Evaluating Capability Elicitation Techniques. *arXiv*, https://arxiv.org/abs/2502.02180

Lucas Hu, Shaown Sarker, Billy Melicher, Alex Starov, Wei Wang, Nabeel Mohamed, and Tony Li (2024) Now You See Me, Now You Don't: Using LLMs to Obfuscate Malicious JavaScript. *Unit 42*, https://unit42.paloaltonetworks.com/using-llms-obfuscate-malicious-javascript/#:~:text=In%202023%2C%20news%20outlets%20published,and%20perform%20other%20malicious%20tasks

Jiaxin Huang, Shixiang Shane Gu, Le Hou, Yuexin Wu, Xuezhi Wang, Hongkun Yu, and Jiawei Han (2022) Large Language Models Can Self-Improve. *arXiv*, https://arxiv.org/abs/2210.11610

Evan Hubinger, Carson Denison, Jesse Mu, Mike Lambert, Meg Tong, Monte MacDiarmid, Tamera Lanham, Daniel M. Ziegler, Tim Maxwell, Newton Cheng, Adam Jermyn, Amanda Askell, Ansh Radhakrishnan, Cem Anil, David Duvenaud, Deep Ganguli, Fazl Barez, Jack Clark, Kamal Ndousse, Kshitij Sachan, Michael Sellitto, Mrinank Sharma, Nova DasSarma, Roger Grosse, Shauna Kravec, Yuntao Bai, Zachary Witten, Marina Favaro, Jan Brauner, Holden Karnofsky, Paul Christiano, Samuel R. Bowman, Logan Graham, Jared Kaplan, Sören Mindermann, Ryan Greenblatt, Buck Shlegeris, Nicholas Schiefer, and Ethan Perez (2024). Sleeper Agents: Training Deceptive LLMs that Persist Through Safety Training. *arXiv*, https://arxiv.org/abs/2401.05566

Evan Hubinger, Chris van Merwijk, Vladimir Mikulik, Joar Skalse, and Scott Garrabrant (2019) Risks from Learned Optimization in Advanced Machine Learning Systems. *arXiv*, https://arxiv.org/abs/1906.01820

Hugging Face (2022) Evaluate. *Hugging Face*, https://huggingface.co/docs/evaluate/index

Humane Intelligence (2024) Generative AI Red Teaming Challenge: Transparency Report. *Humane Intelligence*, https://humane-intelligence.org/wp-content/uploads/2025/09/2024-GenerativeAI-RedTeaming-TransparencyReport.pdf

IAPP (2024) Global AI Law and Policy Tracker. *The International Association of Privacy Professionals*, https://iapp.org/media/pdf/resource_center/global_ai_law_policy_tracker.pdf

IAPP (2025) US State AI Governance Legislation Tracker 2025. *The International Association of Privacy Professionals*, https://iapp.org/media/pdf/resource_center/us_state_ai_governance_legislation_tracker.pdf

IBM (2025) The AI Ethics Trust Engine: Converting Capital into Capabilities. *IBM Institute for Business Value*, https://www.ibm.com/downloads/documents/us-en/1443d5cb56c02cb4

ISO (n.d.) Foreword - Supplementary information. *ISO*, https://www.iso.org/foreword-supplementary-information.html

ISO/IEC (2022) ISO/IEC International Standard 27001:2022, Information security management systems. *ISO*, https://www.iso.org/standard/27001

ITI (2024) Authenticating AI-Generated Content: Exploring Risks, Techniques & Policy Recommendations. *The Information Technology Industry Council (ITI)*, https://www.itic.org/policy/ITI_AIContentAuthorizationPolicy_122123.pdf

Krystal Jackson, Deepika Raman, Jessica Newman, Nada Madkour, Charlotte Yuan, and Evan R. Murphy (2026) Toward Risk Thresholds for AI-Enabled Cyber Threats: Enhancing Decision-Making Under Uncertainty with Bayesian Networks. *arXiv*, https://arxiv.org/abs/2601.17225

Shomik Jain, Vinith Suriyakumar, Kathleen Creel, and Ashia Wilson (2023) Algorithmic Pluralism: A Structural Approach To Equal Opportunity. *arXiv*, https://arxiv.org/abs/2305.08157

Japan AISI (2024) Guide to Red Teaming Methodology on AI Safety (Version 1.00). *Japan AI Safety Institute*, https://aisi.go.jp/assets/pdf/ai_safety_RT_v1.00_en.pdf

Sunny Jiang, Yesilernis Pena, Dell Gines, and Todd Lang (2025) Artificial Intelligence Impact on Labor Markets. International Economic Development Council, https://www.iedconline.org/clientuploads/EDRP%20Logos/AI_Impact_on_Labor_Markets.pdf

Mandar Joshi, Eunsol Choi, Daniel S. Weld, and Luke Zettlemoyer (2017a) TriviaQA: A Large Scale Distantly Supervised Challenge Dataset for Reading Comprehension. *arXiv*, https://arxiv.org/abs/1705.03551

Mandar Joshi, Eunsol Choi, Daniel S. Weld, and Luke Zettlemoyer (2017b) TriviaQA: A Large Scale Distantly Supervised Challenge Dataset for Reading Comprehension. *GitHub*, https://github.com/mandarjoshi90/triviaqa

Mandar Joshi, Eunsol Choi, Daniel S. Weld, and Luke Zettlemoyer (2017c) TriviaQA: A Large Scale Dataset for Reading Comprehension and Question Answering. *TriviaQA*, https://nlp.cs.washington.edu/triviaqa/

Sayash Kapoor, Rishi Bommasani, Kevin Klyman, Shayne Longpre, Ashwin Ramaswami, Peter Cihon, Aspen Hopkins, Kevin Bankston, Stella Biderman, Miranda Bogen, Rumman Chowdhury, Alex Engler, Peter Henderson, Yacine Jernite, Seth Lazar, Stefano Maffulli, Alondra Nelson, Joelle Pineau, Aviya Skowron, Dawn Song, Victor Storchan, Daniel Zhang, Daniel E. Ho, Percy Liang, and Arvind Narayanan (2024) On the Societal Impact of Open Foundation Models. *arXiv*, https://arxiv.org/abs/2403.07918

Holden Karnofsky (2021) Forecasting Transformative AI, Part 1: What Kind of AI? *Cold Takes*, https://www.cold-takes.com/transformative-ai-timelines-part-1-of-4-what-kind-of-ai/

Holden Karnofsky (2024) If-Then Commitments for AI Risk Reduction. *Carnegie Endowment for International Peace*, https://carnegie-production-assets.s3.amazonaws.com/static/files/Karnofsky%20-%20AI%20Risk%20Reduction-1.pdf

Kiran Karra, Chace Ashcraft, and Neil Fendley (2020) The TrojAI Software Framework: An Open Source tool for Embedding Trojans into Deep Learning Models. *arXiv*, https://arxiv.org/abs/2003.07233

Atoosa Kasirzadeh (2024) Measurement Challenges in AI Catastrophic Risk Governance and Safety Frameworks. *Tech Policy Press*, https://www.techpolicy.press/measurement-challenges-in-ai-catastrophic-risk-governance-and-safety-frameworks/

Atoosa Kasirzadeh and Iason Gabriel (2025) Characterizing AI Agents for Alignment and Governance. *arXiv*, https://arxiv.org/pdf/2504.21848

Davinder Kaur, Suleyman Uslu, Kaley J. Rittichier, and Arjan Durresi (2022) Trustworthy Artificial Intelligence: A Review. *ACM Computing Surveys*, https://doi.org/10.1145/3491209

Zachary Kenton, Ramana Kumar, Sebastian Farquhar, Jonathan Richens, Matt MacDermott, and Tom Everitt (2022) Discovering Agents. *arXiv*, https://arxiv.org/abs/2208.08345

Josh Kenway, Camille François, Sasha Costanza-Chock, Inioluwa Deborah Raji, and Joy Buolamwini (2022) Bug Bounties for Algorithmic Harms? *Algorithmic Justice League*, https://www.ajl.org/bugs

Prakash L. Kharvi (2024) Understanding the Impact of AI-Generated Deepfakes on Public Opinion, Political Disclosure, and Personal Security in Social Media. *IEEE Security & Privacy*, https://www.computer.org/csdl/magazine/sp/2024/04/10552098/1XApkaTs5l6

Heidy Khlaaf, Pamela Mishkin, Joshua Achiam, Gretchen Krueger, and Miles Brundage (2022) A Hazard Analysis Framework for Code Synthesis Large Language Models. *arXiv*, https://arxiv.org/abs/2207.14157

Heidy Khlaaf and Sarah Myers West (2025) Safety Co-Option and Compromised National Security: The Self-Fulfilling Prophecy of Weakened AI Risk Thresholds. *arXiv*, https://arxiv.org/pdf/2504.15088

Douwe Kiela, Max Bartolo, Yixin Nie, Divyansh Kaushik, Atticus Geiger, Zhengxuan Wu, Bertie Vidgen, Grusha Prasad, Amanpreet Singh, Pratik Ringshia, Zhiyi Ma, Tristan Thrush, Sebastian Riedel, Zeerak Waseem, Pontus Stenetorp, Robin Jia, Mohit Bansal, Christopher Potts, and Adina Williams (2021) Dynabench: Rethinking Benchmarking in NLP. *arXiv*, https://arxiv.org/abs/2104.14337

Megan Kinniment, Lucas Jun Koba Sato, Haoxing Du, Brian Goodrich, Max Hasin, Lawrence Chan, Luke Harold Miles, Tao R. Lin, Hjalmar Wijk, Joel Burget, Aaron Ho, Elizabeth Barnes, and Paul Christiano (2023) Evaluating Language-Model Agents on Realistic Autonomous Tasks. *Alignment Research Center*, https://evals.alignment.org/Evaluating_LMAs_Realistic_Tasks.pdf

Kevin Klyman (2024) Acceptable Use Policies for Foundation Models. *arXiv*, https://arxiv.org/abs/2409.09041

Leonie Koessler, Jonas Schuett, and Markus Anderljung (2024) Risk thresholds for frontier AI. *arXiv*, https://arxiv.org/abs/2406.14713

Victoria Krakovna, Jonathan Uesato, Vladimir Mikulik, Matthew Rahtz, Tom Everitt, Ramana Kumar, Zac Kenton, Jan Leike, and Shane Legg (2020) Specification gaming: the flip side of AI ingenuity. *Google DeepMind,* https://www.deepmind.com/blog/specification-gaming-the-flip-side-of-ai-ingenuity

Julia Kreutzer, Isaac Caswell, Lisa Wang, Ahsan Wahab, Daan van Esch, Nasanbayar Ulzii-Orshikh, Allahsera Tapo, Nishant Subramani, Artem Sokolov, Claytone Sikasote, Monang Setyawan, Supheakmungkol Sarin, Sokhar Samb, Benoît Sagot, Clara Rivera, Annette Rios, Isabel Papadimitriou, Salomey Osei, Pedro Ortiz Suarez, Iroro Orife, Kelechi Ogueji, Andre Niyongabo Rubungo, Toan Q. Nguyen, Mathias Müller, André Müller, Shamsuddeen Hassan Muhammad, Nanda Muhammad, Ayanda Mnyakeni, Jamshidbek Mirzakhalov, Tapiwanashe Matangira, Colin Leong, Nze Lawson, Sneha Kudugunta, Yacine Jernite, Mathias Jenny, Orhan Firat, Bonaventure F. P. Dossou, Sakhile Dlamini, Nisansa de Silva, Sakine Çabuk Ballı, Stella Biderman, Alessia Battisti, Ahmed Baruwa, Ankur Bapna, Pallavi Baljekar, Israel Abebe Azime, Ayodele Awokoya, Duygu Ataman, Orevaoghene Ahia, Oghenefego Ahia, Sweta Agrawal, and Mofetoluwa Adeyemi (2022) Quality at a Glance: An Audit of Web-Crawled Multilingual Datasets. Transactions of the Association for Computational Linguistics 10 (2022), 50–72. https://doi.org/10.1162/tacl_a_00447

Atul Kumar (2025) Why AI is Replacing Some Jobs Faster Than Others. *World Economic Forum,* https://www.weforum.org/stories/2025/08/ai-jobs-replacement-data-careers/

Alexandre Lacoste, Alexandra Luccioni, Victor Schmidt, and Thomas Dandres (2019) Quantifying the Carbon Emissions of Machine Learning. *arXiv*, https://arxiv.org/abs/1910.09700

Rudolf Laine, Alexander Meinke, and Owain Evans (2023) Towards a Situational Awareness Benchmark for LLMs. *NeurIPS 2023 Workshop, Socially Responsible Language Modelling Research* (*SoLaR*), https://openreview.net/forum?id=DRk4bWKr41

Rudolf Laine, Bilal Chughtai, Jan Betley, Kaivalya Hariharan, Jeremy Scheurer, Mikita Balesni, Marius Hobbhahn, Alexander Meinke, and Owain Evans (2024) Me, Myself, and AI: The Situational Awareness Dataset (SAD) for LLMs. *arXiv*, https://arxiv.org/abs/2407.04694

Lauro Langosco, Jack Koch, Lee Sharkey, Jacob Pfau, Laurent Orseau, and David Krueger (2021) Goal Misgeneralization in Deep Reinforcement Learning. *arXiv*, https://arxiv.org/abs/2105.14111

Pat Langley (2024) Integrated Systems for Computational Scientific Discovery. *Proceedings of the AAAI Conference on Artificial Intelligence, 38(20), 22598–22606*, https://doi.org/10.1609/aaai.v38i20.30269

Gregor von Laszewski, Ben Hawks, Marco Colombo,Reece Shiraishi, Anjay Krishnan, Nhan Tran, and Geoffrey C. Fox (2025) MLCommons Science Working Group AI Benchmarks Collection. *MLCommons,* https://mlcommons-science.github.io/benchmark/benchmarks.pdf

Jon M. Laurent, Joseph D. Janizek, Michael Ruzo, Michaela M. Hinks, Michael J. Hammerling, Siddharth Narayanan, Manvitha Ponnapati, Andrew D. White, and Samuel G. Rodriques (2024) LAB-Bench: Measuring Capabilities of Language Models for Biology Research. *arXiv*, https://arxiv.org/abs/2407.10362

Seanie Lee, Minsu Kim, Lynn Cherif, David Dobre, Juho Lee, Sung Ju Hwang, Kenji Kawaguchi, Gauthier Gidel, Yoshua Bengio, Nikolay Malkin, and Moksh Jain (2025) Learning Diverse Attacks on Large Language Models for Robust Red-Teaming and Safety Tuning. *arXiv*, https://arxiv.org/abs/2405.18540

Brenda Leong and Daniel Atherton (2023) AI incident response plans: Not just for security anymore. *iapp,* https://iapp.org/news/a/ai-incident-response-plans-not-just-for-security-anymore

Simon Lermen, Mateusz Dziemian, and Govind Pimpale (2024) Applying Refusal-Vector Ablation to Llama 3.1 70B Agents. *arXiv*, https://arxiv.org/abs/2410.10871

Yujia Li, David Choi, Junyoung Chung, Nate Kushman, Julian Schrittwieser, Rémi Leblond, Tom Eccles, James Keeling, Felix Gimeno, Agustin Dal Lago, Thomas Hubert, Peter Choy, Cyprien de Masson d'Autume, Igor Babuschkin, Xinyun Chen, Po-Sen Huang, Johannes Welbl, Sven Gowal, Alexey Cherepanov, James Molloy, Daniel J. Mankowitz, Esme Sutherland Robson, Pushmeet Kohli, Nando de Freitas, Koray Kavukcuoglu, and Oriol Vinyals (2022) Competition-Level Code Generation with AlphaCode. *Google DeepMind,* https://storage.googleapis.com/deepmind-media/AlphaCode/competition_level_code_generation_with_alphacode.pdf

Yiming Li, Shuo Shao, Yu He, Junfeng Guo, Tianwei Zhang, Zhan Qin, Pin-Yu Chen, Michael Backes, Philip Torr, Dacheng Tao, and Kui Ren (2025) Rethinking Data Protection in the (Generative) Artificial Intelligence Era. *arXiv*, https://arxiv.org/html/2507.03034v4

Lin Li, Guikun Chen, Hanrong Shi, Jun Xiao, and Long Chen (2024) A Survey on Multimodal Benchmarks: In the Era of Large AI Models. *arXiv*, https://arxiv.org/abs/2409.18142

Nathaniel Li, Alexander Pan, Anjali Gopal, Summer Yue, Daniel Berrios, Alice Gatti, Justin D. Li, Ann-Kathrin Dombrowski, Shashwat Goel, Long Phan, Gabriel Mukobi, Nathan Helm-Burger, Rassin Lababidi, Lennart Justen, Andrew B. Liu, Michael Chen, Isabelle Barrass, Oliver Zhang, Xiaoyuan Zhu, Rishub Tamirisa, Bhrugu Bharathi, Adam Khoja, Ariel Herbert-Voss, Cort B. Breuer, Andy Zou, Mantas Mazeika, Zifan Wang, Palash Oswal, Weiran Liu, Adam A. Hunt, Justin Tienken-Harder, Kevin Y. Shih, Kemper Talley, John Guan, Russell Kaplan, Ian Steneker, David Campbell, Brad Jokubaitis, Alex Levinson, Jean Wang, William Qian, Kallol Krishna Karmakar, Steven Basart, Stephen Fitz, Mindy Levine, Ponnurangam Kumaraguru, Uday Tupakula, Vijay Varadharajan, Yan Shoshitaishvili, Jimmy Ba, Kevin M. Esvelt, Alexandr Wang, and Dan Hendrycks (2024a) The WMDP Benchmark: Measuring and Reducing Malicious Use With Unlearning. *arXiv*, https://arxiv.org/abs/2403.03218

Nathaniel Li, Alexander Pan, Anjali Gopal, Summer Yue, Daniel Berrios, Alice Gatti, Justin D. Li, Ann-Kathrin Dombrowski, Shashwat Goel, Long Phan, Gabriel Mukobi, Nathan Helm-Burger, Rassin Lababidi, Lennart Justen, Andrew B. Liu, Michael Chen, Isabelle Barrass, Oliver Zhang, Xiaoyuan Zhu, Rishub Tamirisa, Bhrugu Bharathi, Adam Khoja, Ariel Herbert-Voss, Cort B. Breuer, Andy Zou, Mantas Mazeika, Zifan Wang, Palash Oswal, Weiran Liu, Adam A. Hunt, Justin Tienken-Harder, Kevin Y. Shih, Kemper Talley, John Guan, Russell Kaplan, Ian Steneker, David Campbell, Brad Jokubaitis, Alex Levinson, Jean Wang, William Qian, Kallol Krishna

Karmakar, Steven Basart, Stephen Fitz, Mindy Levine, Ponnurangam Kumaraguru, Uday Tupakula, Vijay Varadharajan, Yan Shoshitaishvili, Jimmy Ba, Kevin M. Esvelt, Alexandr Wang, and Dan Hendrycks (2024b) The WMDP Benchmark: Measuring and Reducing Malicious Use With Unlearning. *GitHub*, https://github.com/centerforaisafety/wmdp

Nathaniel Li, Alexander Pan, Anjali Gopal, Summer Yue, Daniel Berrios, Alice Gatti, Justin D. Li, Ann-Kathrin Dombrowski, Shashwat Goel, Long Phan, Gabriel Mukobi, Nathan Helm-Burger, Rassin Lababidi, Lennart Justen, Andrew B. Liu, Michael Chen, Isabelle Barrass, Oliver Zhang, Xiaoyuan Zhu, Rishub Tamirisa, Bhrugu Bharathi, Adam Khoja, Ariel Herbert-Voss, Cort B. Breuer, Andy Zou, Mantas Mazeika, Zifan Wang, Palash Oswal, Weiran Liu, Adam A. Hunt, Justin Tienken-Harder, Kevin Y. Shih, Kemper Talley, John Guan, Russell Kaplan, Ian Steneker, David Campbell, Brad Jokubaitis, Alex Levinson, Jean Wang, William Qian, Kallol Krishna Karmakar, Steven Basart, Stephen Fitz, Mindy Levine, Ponnurangam Kumaraguru, Uday Tupakula, Vijay Varadharajan, Yan Shoshitaishvili, Jimmy Ba, Kevin M. Esvelt, Alexandr Wang, and Dan Hendrycks (2024c) Datasets: cais/wmpd. *Hugging Face*, https://huggingface.co/datasets/cais/wmdp

Hui Li, Zhouyang Ren, Miao Fan, Wenyuan Li, Yan Xu, Yunpeng Jiang, and Weiyi Xia (2022) A Review of Scenario Analysis Methods in Planning and Operation of Modern Power Systems: Methodologies, Applications, and Challenges. *ScienceDirect*, https://www.sciencedirect.com/science/article/abs/pii/S0378779621007033?via%3Dihub

Yuanchun Li, Hao Wen, Weijun Wang, Xiangyu Li, Yizhen Yuan, Guohong Liu, Jiacheng Liu, Wenxing Xu, Xiang Wang, Yi Sun, Rui Kong, Yile Wang, Hanfei Geng, Jian Luan, Xuefeng Jin, Zilong Ye, Guanjing Xiong, Fan Zhang, Xiang Li, Mengwei Xu, Zhijun Li, Peng Li, Yang Liu, Ya-Qin Zhang, and Yunxin Liu (2024). Personal LLM Agents: Insights and Survey about the Capability, Efficiency and Security. *arXiv*, https://arxiv.org/abs/2401.05459

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, Benjamin Newman, Binhang Yuan, Bobby Yan, Ce Zhang, Christian Cosgrove, Christopher D. Manning, Christopher Ré, Diana Acosta-Navas, Drew A. Hudson, Eric Zelikman, Esin Durmus, Faisal Ladhak, Frieda Rong, Hongyu Ren, Huaxiu Yao, Jue Wang, Keshav Santhanam, Laurel Orr, Lucia Zheng, Mert Yuksekgonul, Mirac Suzgun, Nathan Kim, Neel Guha, Niladri Chatterji, Omar Khattab, Peter Henderson, Qian Huang, Ryan Chi, Sang Michael Xie, Shibani Santurkar, Surya Ganguli, Tatsunori Hashimoto, Thomas Icard, Tianyi Zhang, Vishrav Chaudhary, William Wang, Xuechen Li, Yifan Mai, Yuhui Zhang, and Yuta Koreeda (2022) Holistic Evaluation of Language Models. *arXiv*, https://arxiv.org/abs/2211.09110

Weixin Liang, Nazneen Rajani, Xinyu Yang, Ezinwanne Ozoani, Eric Wu, Yiqun Chen, Daniel Scott Smith, and James Zou (2024) What's documented in AI? Systematic Analysis of 32K AI Model Cards. *arXiv*, https://arxiv.org/abs/2402.05160

Stephanie Lin, Jacob Hilton, and Owain Evans (2021a) TruthfulQA: Measuring How Models Mimic Human Falsehoods. *arXiv*, https://arxiv.org/abs/2109.07958

Stephanie Lin, Jacob Hilton, and Owain Evans (2021b) TruthfulQA: Measuring How Models Mimic Human Falsehoods. *GitHub*, https://github.com/sylinrl/TruthfulQA

Xiaojian Lin and Michael Losavio (2025) A Comprehensive Survey on Bias and Fairness in Generative AI: Legal, Ethical, and Technical Responses. *SpringerNature*, https://link.springer.com/chapter/10.1007/978-981-96-7273-8_22

Shayne Longpre, Sayash Kapoor, Kevin Klyman, Ashwin Ramaswami, Rishi Bommasani, Borhane Blili-Hamelin, Yangsibo Huang, Aviya Skowron, Zheng-Xin Yong, Suhas Kotha, Yi Zeng, Weiyan Shi, Xianjun Yang, Reid Southen, Alexander Robey, Patrick Chao, Diyi Yang, Ruoxi Jia, Daniel Kang, Sandy Pentland, Arvind Narayanan, Percy Liang, and Peter Henderson (2024) A Safe Harbor for AI Evaluation and Red Teaming. *arXiv*, https://arxiv.org/abs/2403.04893

Shayne Longpre, Kevin Klyman, Ruth E. Appel, Sayash Kapoor, Rishi Bommasani, Michelle Sahar, Sean McGregor, Avijit Ghosh, Borhane Blili-Hamelin, Nathan Butters, Alondra Nelson, Amit Elazari, Andrew Sellars, Casey John Ellis, Dane Sherrets, Dawn Song, Harley Geiger, Ilona Cohen, Lauren McIlvenny, Madhulika Srikumar, Mark M. Jaycox, Markus Anderljung, Nadine Farid Johnson, Nicholas Carlini, Nicolas Miailhe, Nik Marda, Peter Henderson, Rebecca S. Portnoff, Rebecca Weiss, Victoria Westerhoff, Yacine Jernite, Rumman Chowdhury, Percy Liang, and Arvind Narayanan (2025) In-House Evaluation Is Not Enough: Towards Robust Third-Party Flaw Disclosure for General-Purpose AI. *arXiv*, https://arxiv.org/abs/2503.16861

Gary D. Lopez Munoz, Amanda J. Minnich, Roman Lutz, Richard Lundeen, Raja Sekhar Rao Dheekonda, Nina Chikanov, Bolor-Erdene Jagdagdorj, Martin Pouliot, Shiven Chawla, Whitney Maxwell, Blake Bullwinkel, Katherine Pratt, Joris de Gruyter, Charlotte Siska, Pete Bryan, Tori Westerhoff, Chang Kawaguchi, Christian Seifert, Ram Shankar Siva Kumar, and Yonatan Zunger (2024) PyRIT: A Framework for Security Risk Identification and Red Teaming in Generative AI System. *arXiv*, https://arxiv.org/abs/2410.02828v1

Chris Lu, Cong Lu, Robert Tjarko Lange, Jakob Foerster, Jeff Clune, and David Ha (2024) The AI Scientist: Towards Fully Automated Open-Ended Scientific Discovery. *arXiv*, https://www.arxiv.org/abs/2408.06292

Alexandra Sasha Luccioni, Sylvain Viguier, and Anne-Laure Ligozat (2022). Estimating the Carbon Footprint of BLOOM, a 176B Parameter Language Model. *arXiv*, https://arxiv.org/abs/2211.02001

Alexandra Sasha Luccioni and Alex Hernandez-Garcia (2023) Counting Carbon: A Survey of Factors Influencing the Emissions of Machine Learning. *arXiv*, https://arxiv.org/abs/2302.08476

Linhao Luo, Yuan-Fang Li, Gholamreza Haffari, and Shirui Pan (2024) Reasoning on Graphs: Faithful and Interpretable Large Language Model Reasoning. *ICLR*, https://openreview.net/pdf?id=ZGNWW7xZ6Q

Matthijs M. Maas (2022) Aligning AI Regulation to Sociotechnical Change. *Oxford Academic*, https://academic.oup.com/edited-volume/41989/chapter/355438659

Nada Madkour, Jessica Newman, Deepika Raman, Krystal Jackson, Evan R. Murphy, and Charlotte Yuan (2026a) Agentic AI Risk-Management Standards Profile. *UC Berkeley Center for Long-Term Cybersecurity,* https://cltc.berkeley.edu/wp-content/uploads/2026/02/Agentic-AI-Risk-Management-Standards-Profile.pdf

Nada Madkour, Jessica Newman, Deepika Raman, Krystal Jackson, Evan R. Murphy, and Charlotte Yuan (2026b) Transparency, Documentation, and Reporting Recommendations for General-Purpose AI Risk-Management. *UC Berkeley Center for Long-Term Cybersecurity,* https://cltc.berkeley.edu/wp-content/uploads/2026/04/Berkeley-Transparency-Recommendations-for-GPAI-Risk-Management.pdf

Nada Madkour, Jessica Newman, Deepika Raman, Krystal Jackson, Evan R. Murphy, and Charlotte Yuan (2026c) Mapping Key Standards and Regulations to the General-Purpose AI Risk Management Standards Profile V1.2 *UC Berkeley Center for Long-Term Cybersecurity,* https://cltc.berkeley.edu/wp-content/uploads/2026/04/Berkeley-Mapping-of-Profile-Guidance-v1-2-to-Key-Standards-and-Regulations.pdf

Nada Madkour, Jessica Newman, Deepika Raman, Krystal Jackson, Evan R. Murphy, and Charlotte Yuan (2026d) Evaluation of Frontier AI Company Practices Using the General-Purpose AI Risk-Management Standards Profile V1.2. *UC Berkeley Center for Long-Term Cybersecurity,* https://cltc.berkeley.edu/wp-content/uploads/2026/04/Berkeley-Evaluation-of-Frontier-AI-v1-2.pdf

Mehrdad Maghsoudi, Amirmahdi Mohammadi, and Sajjad Habibipour (2024) Navigating and Addressing Public Concerns in AI: Insights From Social Media Analytics and Delphi. *IEEE*, https://ieeexplore.ieee.org/document/10630673

Alessandro Mantelero (2024) The Fundamental Rights Impact Assessment (FRIA) in the AI Act: Roots, legal obligations and key elements for a model template. *ScienceDirect*, https://www.sciencedirect.com/science/article/pii/S0267364924000864?via%3Dihub

Fernando Martínez-Plumed and José Hernández-Orallo (2020) Dual Indicators to Analyze AI Benchmarks: Difficulty, Discrimination, Ability, and Generality. *IEEE Transactions on Games* 12 (2) pp. 121–131, June 2020, doi:10.1109/TG.2018.2883773. https://ieeexplore.ieee.org/document/8550672

Timothy R. McIntosh, Teo Susnjak, Nalin Arachchilage, Tong Liu, Paul Watters, and Malka N. Halgamuge (2024) Inadequacies of Large Language Model Benchmarks in the Era of Generative Artificial Intelligence. *arXiv*, https://arxiv.org/abs/2402.09880

Alexander Meinke, Bronson Schoen, Jérémy Scheurer, Mikita Balesni, Rusheb Shah, and Marius Hobbhahn (2025) Frontier Models are Capable of In-context Scheming. *arXiv*, https://arxiv.org/abs/2412.04984

Meta AI (2024a) Llama Responsible Use Guide. *Meta*, https://ai.meta.com/static-resource/july-responsible-use-guide

Meta AI (2024b) Llama 3.1 Acceptable Use Policy. *Meta*, https://llama.meta.com/llama3_1/use-policy/

Meta AI (2025a) Llama 4 Acceptable Use Policy. *Meta*, https://www.llama.com/llama4/use-policy/

Meta AI (2025b) Frontier AI Framework. *Meta*, https://about.fb.com/news/2025/02/meta-approach-frontier-ai/

METR (n.d.a) Frontier AI Safety Policies. *METR*, https://metr.org/faisc

METR (n.d.b) Key Components of an RSP. *METR*, https://metr.org/rsp-key-components/

METR (2024) METR's Autonomy Evaluation Resources. *Model Evaluation and Threat Research*, https://metr.github.io/autonomy-evals-guide/

METR (2025) Common Elements of Frontier AI Safety Policies. *METR*, https://metr.org/blog/2025-03-26-common-elements-of-frontier-ai-safety-policies/

Microsoft (n.d.) microsoft/autogen. *GitHub*, https://github.com/microsoft/autogen

Microsoft (2022a) Responsible use of AI with Cognitive Services. *Microsoft,* https://docs.microsoft.com/en-us/azure/cognitive-services/responsible-use-of-ai-overview

Microsoft (2022b) Responsible AI Impact Assessment Guide. *Microsoft,* https://blogs.microsoft.com/wp-content/uploads/prod/sites/5/2022/06/Microsoft-RAI-Impact-Assessment-Guide.pdf

Microsoft (2023) Governing AI: A Blueprint for the Future. *Microsoft,* https://cdn-dynmedia-1.microsoft.com/is/content/microsoftcorp/microsoft/msc/documents/presentations/CSR/Governing-AI-A-Blueprint-for-the-Future.pdf

Microsoft (2025) Frontier Governance Framework. *Microsoft,* https://cdn-dynmedia-1.microsoft.com/is/content/microsoftcorp/microsoft/final/en-us/microsoft-brand/documents/Microsoft-Frontier-Governance-Framework.pdf

Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru (2019) Model cards for model reporting. In *Proceedings of the conference on fairness, accountability, and transparency* 2019, pp. 220–229, https://dl.acm.org/doi/10.1145/3287560.3287596

Margaret Mitchell, Avijit Ghosh, Alexandra Sasha Luccioni, and Giada Pistilli (2025) Fully Autonomous AI Agents Should Not be Developed. *arXiv*, https://arxiv.org/abs/2502.02649

MIT (n.d.a) AI Risk Repository. *Massachusetts Institute of Technology,* https://airisk.mit.edu/

MIT (n.d.b) AI Incident Tracker. *Massachusetts Institute of Technology,* https://airisk.mit.edu/ai-incident-tracker

MIT (n.d.c) How can we reduce the risks of AI? A Comprehensive Living Database of Over 800 AI Risk Mitigations and a Preliminary Taxonomy. *Massachusetts Institute of Technology,* https://airisk.mit.edu/ai-risk-mitigations

MITRE (n.d.a) Navigate Threats to AI Systems Through Real-World Insights. *MITRE,* https://atlas.mitre.org/

MITRE (n.d.b) ATLAS AI Incidents: Submit & Receive Anonymized Community AI Incident Data. *MITRE,* https://ai-incidents.mitre.org/

Nicolas Moës, Toni Lorente, and Yolanda Lannquist (2023) List of Potential Clauses to Govern the Development of General Purpose AI Systems (GPAIS): Draft Version 0.1. *The Future Society,* http://thefuturesociety.org/potential-clauses-to-govern-gpais

Jakob Mökander, Jonas Schuett, Hannah Rose Kirk, and Luciano Floridi (2023) Auditing large language models: a three-layered approach. *AI and Ethics,* https://doi.org/10.1007/s43681-023-00289-2

M. Granger Morgan and Max Henrion (1990) Uncertainty: A Guide to Dealing with Uncertainty in Quantitative Risk and Policy Analysis. *Cambridge University Press, New York.* https://doi.org/10.1017/CBO9780511840609

M. Granger Morgan (2017) Theory and Practice in Policy Analysis. *Cambridge University Press, New York,* https://doi.org/10.1017/9781316882665

Christopher A. Mouton, Caleb Lucas, and Ella Guest (2024) The Operational Risks of AI in Large-Scale Biological Attacks: Results of a Red-Team Study. *RAND Corporation*, https://www.rand.org/pubs/research_reports/RRA2977-2.html

Gabriel Mukobi (2024) Reasons to Doubt the Impact of AI Risk Evaluations. *arXiv*, https://arxiv.org/abs/2408.02565

Mark Muro, Jacob Whiton, and Robert Maxim (2019) What Jobs Are Affected By AI? *Brookings Artificial Intelligence and Emerging Technology Initiative,* https://www.brookings.edu/articles/what-jobs-are-affected-by-ai-better-paid-better-educated-workers-face-the-most-exposure/

Malcolm Murray, Henry Papadatos, Otter Quarks, Pierre-François Gimenez, Simeon Campos (2025) Mapping AI Benchmark Data to Quantitative Risk Estimates Through Expert Elicitation. *arXiv*, https://arxiv.org/pdf/2503.04299

Mekhail Mustak, Joni Salminen, Matti Mäntymäki, Arafat Rahman, and Yogesh K. Dwivedi (2023) Deepfakes: Deceptions, Mitigations, and Opportunities. *ScienceDirect,* https://www.sciencedirect.com/science/article/abs/pii/S0148296322008335

Reiichiro Nakano, Jacob Hilton, Suchir Balaji, Jeff Wu, Long Ouyang, Christina Kim, Christopher Hesse, Shantanu Jain, Vineet Kosaraju, William Saunders, Xu Jiang, Karl Cobbe, Tyna Eloundou, Gretchen Krueger, Kevin Button, Matthew Knight, Benjamin Chess, and John Schulman (2021) WebGPT: Browser-assisted question-answering with human feedback. *arXiv*, https://arxiv.org/pdf/2112.09332.pdf

Sella Nevo, Dan Lahav, Ajay Karpur, Yogev Bar-On, Henry Alexander Bradley, and Jeff Alstott (2024) Securing AI Model Weights: Preventing Theft and Misuse of Frontier Models. RR-A2849-1. *RAND Corporation,* https://www.rand.org/content/dam/rand/pubs/research_reports/RRA2800/RRA2849-1/RAND_RRA2849-1.pdf

New York State Senate (2026) Responsible AI Safety and Education Act (RAISE Act). *New York State Senate,* https://legislation.nysenate.gov/pdf/bills/2025/S8828

Jessica Newman (2023) A Taxonomy of Trustworthiness for Artificial Intelligence: Connecting Properties of Trustworthiness with Risk-Management and the AI Lifecycle. *UC Berkeley Center for Long-Term Cybersecurity,* https://cltc.berkeley.edu/wp-content/uploads/2023/01/Taxonomy_of_AI_Trustworthiness.pdf

Helen Ngo, Tristan Thrush, Abhishek Thakur, Lewis Tunstall, and Douwe Kiela (2022) Very Large Language Models and How to Evaluate Them. *Hugging Face,* https://huggingface.co/blog/zero-shot-eval-on-the-hub

Richard Ngo, Lawrence Chan, and Sören Mindermann (2022) The alignment problem from a deep learning perspective. *arXiv*, https://arxiv.org/abs/2209.00626

NIST (n.d.a) TrojAI Test and Evaluation Documentation. *National Institute of Standards and Technology,* https://pages.nist.gov/trojai/docs/index.html

NIST (n.d.b) The Language of Trustworthy AI: An In-Depth Glossary of Terms. *National Institute of Standards and Technology,* https://airc.nist.gov/AI_RMF_Knowledge_Base/Glossary

NIST (n.d.c) National Vulnerability Database. *National Institute of Standards and Technology,* https://nvd.nist.gov/

NIST (n.d.d) NIST AI Consortium. *National Institute of Standards and Technology,* https://www.nist.gov/artificial-intelligence/nist-ai-consortium

NIST (2020a) Control Baselines for Information Systems and Organizations. Special Publication 800-53B. *National Institute of Standards and Technology,* https://csrc.nist.gov/pubs/sp/800/53/b/upd1/final

NIST (2020b) Protecting Controlled Unclassified Information in Nonfederal Systems and Organizations. Special Publication 800-171 Rev. 2. *National Institute of Standards and Technology,* https://csrc.nist.gov/pubs/sp/800/171/r2/upd1/final

NIST (2020c) NIST Privacy Framework and Cybersecurity Framework to NIST Special Publication 800-53, Revision 5 Crosswalk. *National Institute of Standards and Technology,* https://www.nist.gov/privacy-framework/nist-privacy-framework-and-cybersecurity-framework-nist-special-publication-800-53

NIST (2021) Enhanced Security Requirements for Protecting Controlled Unclassified Information: A Supplement to NIST Special Publication 800-171. *National Institute of Standards and Technology,* https://csrc.nist.gov/pubs/sp/800/172/final

NIST (2023a) Artificial Intelligence Risk-Management Framework (AI RMF 1.0). AI 100-1. *National Institute of Standards and Technology,* https://doi.org/10.6028/NIST.AI.100-1

NIST (2023b) AI Risk-Management Framework Playbook (version released January 2023). *National Institute of Standards and Technology,* https://www.nist.gov/itl/ai-risk-management-framework/nist-ai-rmf-playbook

NIST (2023c) Biden-Harris Administration Announces New NIST Public Working Group on AI. *National Institute of Standards and Technology,* https://www.nist.gov/news-events/news/2023/06/biden-harris-administration-announces-new-nist-public-working-group-ai

NIST (2023d) NIST SP 800-53, Revision 5 Control Mappings to ISO/IEC 27001. *National Institute of Standards and Technology,* https://csrc.nist.rip/csrc/media/publications/sp/800-53/rev-5/final/documents/sp800-53r5-to-iso-27001-mapping.docx

NIST (2024a) The NIST Cybersecurity Framework (CSF) 2.0. *National Institute of Standards and Technology,* https://nvlpubs.nist.gov/nistpubs/CSWP/NIST.CSWP.29.pdf

NIST (2024b) Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile. *National Institute of Standards and Technology,* https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.600-1.pdf

NIST (2025) Managing Misuse Risk for Dual-Use Foundation Models, Second Public Draft. NIST AI 800-1 2pd. *National Institute of Standards and Technology,* https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.800-1.ipd2.pdf

Curtis G Northcutt, Anish Athalye, and Jonas Mueller (2021) Pervasive Label Errors in Test Sets Destabilize Machine Learning Benchmarks. *OpenReview,* https://openreview.net/forum?id=XccDXrDNLek

Claudio Novelli, Federico Casolari, Antonino Rotolo, Mariarosaria Taddeo, and Luciano Floridi (2024) AI Risk Assessment: A Scenario-Based, Proportional Methodology for the AI Act. *Springer Nature,* https://link.springer.com/article/10.1007/s44206-024-00095-1

NTIA (2024a) Artificial Intelligence Accountability Policy Report. *National Telecommunications and Information Administration,* https://www.ntia.gov/sites/default/files/publications/ntia-ai-report-final.pdf

NTIA (2024b) Dual-Use Foundation Models with Widely Available Model Weights. *National Telecommunications and Information Administration,* https://www.ntia.gov/sites/default/files/publications/ntia-ai-open-model-report.pdf

NTIA (2024c) AI System Documentation. *National Telecommunications and Information Administration,* https://www.ntia.gov/issues/artificial-intelligence/ai-accountability-policy-report/developing-accountability-inputs-a-deeper-dive/information-flow/ai-system-documentation

Joe O'Brien (2024) Coordinated Disclosure of Dual-Use Capabilities: An Early Warning System for Advanced AI. *Institute for AI Policy and Strategy,* https://www.iaps.ai/research/coordinated-disclosure

OECD (2019) OECD AI Principles Overview. *Organization for Economic Co-operation and Development,* https://oecd.ai/en/ai-principles

OECD (2022a) OECD Framework for the Classification of AI Systems. OECD Digital Economy Papers, No. 323. *Organization for Economic Co-operation and Development,* https://doi.org/10.1787/cb6d9eca-en

OECD (2022b) Measuring the environmental impacts of artificial intelligence compute and applications: The AI footprint. OECD Digital Economy Papers, No. 341. *Organization for Economic Co-operation and Development,* https://doi.org/10.1787/7babf571-en

OECD (2023) Advancing accountability in AI: Governing and managing risks throughout the lifecycle for trustworthy AI. *OECD Digital Economy Papers, No. 349, OECD Publishing, Paris,* https://doi.org/10.1787/2448f04b-en

OECD (2025) Introducing the OECD AI Capability Indicators. *OECD Publishing,* https://doi.org/10.1787/be745f04-en

OECD.AI (n.d.a) The OECD.AI Policy Navigator. *Organization for Economic Co-operation and Development,* https://oecd.ai/en/dashboards/overview

OECD.AI (n.d.b) AIM: AI Incidents and Hazards Monitor. *Organization for Economic Co-operation and Development,* https://oecd.ai/en/incidents

OECD.AI (2025) G7 Reporting Framework — Hiroshima AI Process (HAIP) international code of conduct for organizations developing advanced AI systems. *OECD Policy Observatory,* https://transparency.oecd.ai/

Utkarsh Ojha, Yuheng Li, and Yong Jae Lee (2023) Towards universal fake image detectors that generalize across generative models. *arXiv,* https://arxiv.org/abs/2302.10174

OpenAI (2019a) Better Language Models and Their Implications. *OpenAI,* https://openai.com/blog/better-language-models/

OpenAI (2019b) GPT-2: 6-Month Follow-Up. *OpenAI,* https://openai.com/blog/gpt-2-6-month-follow-up/

OpenAI (2019c) Safety Gym. *OpenAI,* https://openai.com/blog/safety-gym/

OpenAI (2019d) safety-gym. *OpenAI,* https://github.com/openai/safety-gym

OpenAI (2023a) GPT-4 System Card. *OpenAI,* https://cdn.openai.com/papers/gpt-4-system-card.pdf

OpenAI (2023b) GPT-4 Technical Report. *arXiv,* https://arxiv.org/abs/2303.08774

OpenAI (2023c) Announcing OpenAI's Bug Bounty Program. *OpenAI,* https://openai.com/blog/bug-bounty-program

OpenAI (2023d) Preparedness Framework (Beta). *OpenAI,* https://cdn.openai.com/openai-preparedness-framework-beta.pdf

OpenAI (2024b) OpenAI o1 System Card. *OpenAI,* https://cdn.openai.com/o1-system-card-20240917.pdf

OpenAI (2024c) Terms of Use. *OpenAI,* https://openai.com/policies/terms-of-use/

OpenAI (2024d) Building an Early Warning System for LLM-Aided Biological Threat Creation. *OpenAI,* https://openai.com/index/building-an-early-warning-system-for-llm-aided-biological-threat-creation/

OpenAI (2024e) Advancing Red Teaming with People and AI. *OpenAI,* https://openai.com/index/advancing-red-teaming-with-people-and-ai/

OpenAI (2025a) Usage Policies. *OpenAI,* https://openai.com/policies/usage-policies/

OpenAI (2025b) Preparedness Framework, Version 2. *OpenAI,* https://cdn.openai.com/pdf/18a02b5d-6b67-4cec-ab64-68cdfbddebcd/preparedness-framework-v2.pdf

Open Philanthropy (2021) Request for proposals for projects in AI alignment that work with deep learning systems. *Open Philanthropy,* https://www.alignmentforum.org/s/Tp3ryR4AxY56ctGh2/p/H5iePjNKaaYQyZpgR

Laurent Orseau and Stuart Armstrong (2016) Safely Interruptible Agents. *Google DeepMind,* https://intelligence.org/files/Interruptibility.pdf

Nicholas G. Otis, Solene Delecourt, Katelyn Cranney, and Rem Koning (2025) Global Evidence on Gender Gaps and Generative AI. *Center for Open Science,* https://osf.io/preprints/osf/h6a7c_v1

Amin Oueslati and Robin Staes-Polet (2025) Ahead of the Curve: Governing AI Agents Under the EU AI Act. *The Future Society,* https://thefuturesociety.org/wp-content/uploads/2023/04/Report-Ahead-of-the-Curve-Governing-AI-Agents-Under-the-EU-AI-Act-4-June-2025.pdf

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe (2022) Training language models to follow instructions with human feedback. *arXiv*, https://arxiv.org/abs/2203.02155

OWASP (2023a) OWASP Top 10 for Large Language Model Applications. Open Worldwide Application Security Project, https://owasp.org/www-project-top-10-for-large-language-model-applications/

OWASP (2023b) OWASP Machine Learning Security Top Ten. *OWASP,* https://owasp.org/www-project-machine-learning-security-top-10/

OWASP (2025a) OWASP AI Vulnerability Scoring System (AIVSS). *OWASP,* https://aivss.owasp.org/

OWASP (2025b) OWASP Top 10 for LLM Applications 2025. *OWASP,* https://owasp.org/www-project-top-10-for-large-language-model-applications/assets/PDF/OWASP-Top-10-for-LLMs-v2025.pdf

OWASP (2025c) GenAI Red Teaming Guide. *OWASP,* https://genai.owasp.org/resource/genai-red-teaming-guide/

Lorenzo Pacchiardi, Alex J. Chan, Sören Mindermann, Ilan Moscovitz, Alexa Y. Pan, Yarin Gal, Owain Evans, and Jan Brauner (2023) How to Catch an AI Liar: Lie Detection in Black-Box LLMs by Asking Unrelated Questions. *arXiv*, https://arxiv.org/abs/2309.15840

PAI (2022) Publication Norms for Responsible AI. *Partnership on AI,* https://partnershiponai.org/workstream/publication-norms-for-responsible-ai/

PAI (2023a) PAI's Responsible Practices for Synthetic Media: A Framework for Collective Action. *Partnership on AI,* https://syntheticmedia.partnershiponai.org/

PAI (2023b) Responsible Generative AI. Let's get started . . . . *Partnership on AI,* https://partnershiponai.org/responsible-generative-ai-lets-get-started/

PAI (2023c) PAI's Guidance for Safe Foundation Model Deployment: A Framework for Collective Action. *Partnership on AI,* https://partnershiponai.org/modeldeployment/

PAI (2023d) Building a Glossary for Synthetic Media Transparency Methods, Part 1: Indirect Disclosure. *Partnership on AI,* https://partnershiponai.org/glossary-for-synthetic-media-transparency-methods-part-1-indirect-disclosure/

PAI (2024) Stakeholder Engagement for Responsible AI: Introducing PAI's Guidelines for Participatory and Inclusive AI. *Partnership on AI,* https://partnershiponai.org/stakeholder-engagement-for-responsible-ai-introducing-pais-guidelines-for-participatory-and-inclusive-ai/

Alexander Pan, Jun Shern Chan, Andy Zou, Nathaniel Li, Steven Basart, Thomas Woodside, Jonathan Ng, Hanlin Zhang, Scott Emmons, and Dan Hendrycks (2023) Do the Rewards Justify the Means? Measuring Trade-Offs Between Rewards and Ethical Behavior in the MACHIAVELLI Benchmark. *arXiv*, https://arxiv.org/abs/2304.03279

Shirui Pan, Linhao Luo, Yufei Wang, Chen Chen, Jiapu Wang, and Xindong Wu (2024) Unifying Large Language Models and Knowledge Graphs: A Roadmap. *IEEE,* https://ieeexplore.ieee.org/document/10387715

Guansong Pang, Chunhua Shen, Longbing Cao, and Anton Van Den Hengel (2021) Deep Learning for Anomaly Detection: A Review. *ACM Digital Library,* https://dl.acm.org/doi/10.1145/3439950

Denis Paperno, Germán Kruszewski, Angeliki Lazaridou, Quan Ngoc Pham, Raffaella Bernardi, Sandro Pezzelle, Marco Baroni, Gemma Boleda, and Raquel Fernández (2016) The LAMBADA dataset: Word prediction requiring a broad discourse context. *arXiv*, https://arxiv.org/abs/1606.06031

Peter S. Park, Simon Goldstein, Aidan O'Gara, Michael Chen, and Dan Hendrycks (2023) AI Deception: A Survey of Examples, Risks, and Potential Solutions. *arXiv*, https://arxiv.org/abs/2308.14752

Seyeon Park and Xiaoli Nan (2025) Generative AI and misinformation: a scoping review of the role of generative AI in the generation, detection, mitigation, and impact of misinformation. *Springer Nature,* https://link.springer.com/article/10.1007/s00146-025-02620-3?utm

Alicia Parrish, Angelica Chen, Nikita Nangia, Vishakh Padmakumar, Jason Phang, Jana Thompson, Phu Mon Htut, and Samuel R. Bowman (2021a) BBQ: A Hand-Built Bias Benchmark for Question Answering. *arXiv*, https://arxiv.org/abs/2110.08193

Alicia Parrish, Angelica Chen, Nikita Nangia, Vishakh Padmakumar, Jason Phang, Jana Thompson, Phu Mon Htut, and Samuel R. Bowman (2021b) BBQ: Repository for the Bias Benchmark for QA dataset. *GitHub*, https://github.com/nyu-mll/BBQ

Will Pearce and Joseph Lucas (2023) NVIDIA AI Red Team: An Introduction. *NVIDIA,* https://developer.nvidia.com/blog/nvidia-ai-red-team-an-introduction/

Pierre Peigné-Lefebvre, Quentin Feuillade-Montixi, Tom David, and Nicolas Miailhe (2025) LLM Robustness Leaderboard v1-Technical Report. *arXiv,* https://arxiv.org/html/2508.06296v1

Ethan Perez, Saffron Huang, Francis Song, Trevor Cai, Roman Ring, John Aslanides, Amelia Glaese, Nat McAleese, and Geoffrey Irving (2022) Red Teaming Language Models with Language Models. *arXiv*, https://arxiv.org/abs/2202.03286

Ethan Perez, Sam Ringer, Kamilė Lukošiūtė, Karina Nguyen, Edwin Chen, Scott Heiner, Craig Pettit, Catherine Olsson, Sandipan Kundu, Saurav Kadavath, Andy Jones, Anna Chen, Ben Mann, Brian Israel, Bryan Seethor, Cameron McKinnon, Christopher Olah, Da Yan, Daniela Amodei, Dario Amodei, Dawn Drain, Dustin Li, Eli Tran-Johnson, Guro Khundadze, Jackson Kernion, James Landis, Jamie Kerr, Jared Mueller, Jeeyoon Hyun, Joshua Landau, Kamal Ndousse, Landon Goldberg, Liane Lovitt, Martin Lucas, Michael Sellitto, Miranda Zhang, Neerav Kingsland, Nelson Elhage, Nicholas Joseph, Noemí Mercado, Nova DasSarma, Oliver Rausch, Robin Larson, Sam McCandlish, Scott Johnston, Shauna Kravec, Sheer El Showk, Tamera Lanham, Timothy Telleen-Lawton, Tom Brown, Tom Henighan, Tristan Hume, Yuntao Bai, Zac Hatfield-Dodds, Jack Clark, Samuel R. Bowman, Amanda Askell, Roger Grosse, Danny Hernandez, Deep Ganguli, Evan Hubinger, Nicholas Schiefer, and Jared Kaplan (2022a) Discovering Language Model Behaviors with Model-Written Evaluations. *arXiv*, https://arxiv.org/abs/2212.09251

Ethan Perez, Sam Ringer, Kamilė Lukošiūtė, Karina Nguyen, Edwin Chen, Scott Heiner, Craig Pettit, Catherine Olsson, Sandipan Kundu, Saurav Kadavath, Andy Jones, Anna Chen, Ben Mann, Brian Israel, Bryan Seethor, Cameron McKinnon, Christopher Olah, Da Yan, Daniela Amodei, Dario Amodei, Dawn Drain, Dustin Li, Eli Tran-Johnson, Guro Khundadze, Jackson Kernion, James Landis, Jamie Kerr, Jared Mueller, Jeeyoon Hyun, Joshua Landau, Kamal Ndousse, Landon Goldberg, Liane Lovitt, Martin Lucas, Michael Sellitto, Miranda Zhang, Neerav Kingsland, Nelson Elhage, Nicholas Joseph, Noemí Mercado, Nova DasSarma, Oliver Rausch, Robin Larson, Sam McCandlish, Scott Johnston, Shauna Kravec, Sheer El Showk, Tamera Lanham, Timothy Telleen-Lawton, Tom Brown, Tom Henighan, Tristan Hume, Yuntao Bai, Zac Hatfield-Dodds, Jack Clark, Samuel R. Bowman, Amanda Askell, Roger Grosse, Danny Hernandez, Deep Ganguli, Evan Hubinger, Nicholas Schiefer, and Jared Kaplan (2022b) Model-Written Evaluation Datasets. *GitHub*, https://github.com/anthropics/evals

Long Phan, Alice Gatti, Ziwen Han, Nathaniel Li, Josephina Hu, Hugh Zhang, Chen Bo Calvin Zhang, Mohamed Shaaban, John Ling, Sean Shi, Michael Choi, Anish Agrawal, Arnav Chopra, Adam Khoja, Ryan Kim, Richard Ren, Jason Hausenloy, Oliver Zhang, Mantas Mazeika, Dmitry Dodonov, Tung Nguyen, Jaeho Lee, Daron Anderson, Mikhail Doroshenko, Alun Cennyth Stokes, Mobeen Mahmood, Oleksandr Pokutnyi, Oleg Iskra, Jessica P. Wang, John-Clark Levin, Mstyslav Kazakov, Fiona Feng, Steven Y. Feng, Haoran Zhao, Michael Yu, Varun Gangal, Chelsea Zou, Zihan Wang, Serguei Popov, Robert Gerbicz, Geoff Galgon, Johannes Schmitt, Will Yeadon, Yongki Lee, Scott Sauers, Alvaro Sanchez, Fabian Giska, Marc Roth, Søren Riis, Saiteja Utpala, Noah Burns, Gashaw M. Goshu, Mohinder Maheshbhai Naiya, Chidozie Agu, Zachary Giboney, Antrell Cheatom, Francesco Fournier-Facio, Sarah-Jane Crowson, Lennart Finke, Zerui Cheng, Jennifer Zampese, Ryan G. Hoerr, Mark Nandor, Hyunwoo Park, Tim Gehrunger, Jiaqi Cai, Ben McCarty, Alexis C Garretson, Edwin Taylor, Damien Sileo, Qiuyu Ren, Usman Qazi, Lianghui Li, Jungbae Nam, John B. Wydallis, Pavel Arkhipov, Jack Wei Lun Shi, Aras Bacho, Chris G. Willcocks, Hangrui Cao, Sumeet Motwani, Emily de Oliveira Santos, Johannes Veith, Edward Vendrow, Doru Cojoc, Kengo Zenitani, Joshua Robinson, Longke Tang, Yuqi Li, Joshua Vendrow, Natanael Wildner Fraga, Vladyslav Kuchkin, Andrey Pupasov Maksimov, Pierre Marion, Denis Efremov, Jayson Lynch, Kaiqu Liang, Aleksandar Mikov, Andrew Gritsevskiy, Julien Guillod et al. (1010 additional authors not shown) (2025) Humanity's Last Exam. *arXiv,* https://arxiv.org/abs/2501.14249

Mary Phuong, Matthew Aitchison, Elliot Catt, Sarah Cogan, Alexandre Kaskasoli, Victoria Krakovna, David Lindner, Matthew Rahtz, Yannis Assael, Sarah Hodkinson, Heidi Howard, Tom Lieberum, Ramana Kumar, Maria Abi Raad, Albert Webson, Lewis Ho, Sharon Lin, Sebastian Farquhar, Marcus Hutter, Grégoire Delétang, Anian Ruoss, Seliem El-Sayed, Sasha Brown, Anca Dragan, Rohin Shah, Allan Dafoe, and Toby Shevlane (2024) Evaluating Frontier Models for Dangerous Capabilities. *arXiv*, https://arxiv.org/abs/2403.13793

Kelsey Piper (2023) How to test what an AI model can — and shouldn't — do. *Vox,* https://www.vox.com/future-perfect/2023/3/29/23661633/gpt-4-openai-alignment-research-center-open-philanthropy-ai-safety

Thomas Ploug and Søren Holm (2022) Right to Contest AI Diagnostics. *Springer Nature,* https://link.springer.com/rwe/10.1007/978-3-030-64573-1_267

PMI (2017) Guide to the Project Management Body of Knowledge. Sixth Edition. *Project Management Institute, Newtown Square, PA*

Promptfoo (2025) Promptfoo: LLM Evals & Red Teaming. *GitHub*, https://github.com/promptfoo/promptfoo?tab=readme-ov-file

Protect AI (2025) ModelScan: Protection Against Model Serialization Attacks. *GitHub*, https://github.com/protectai/modelscan

Xiangyu Qi, Ruoxi Jia, Yi Zeng, Prateek Mittal, Tinghao Xie, and Pin-Yu Chen (2023) Fine-Tuning Aligned Language Models Compromises Safety, Even When Users Do Not Intend to! *arXiv*, https://arxiv.org/pdf/2310.03693

Jack W. Rae, Sebastian Borgeaud, Trevor Cai, Katie Millican, Jordan Hoffmann, Francis Song, John Aslanides, Sarah Henderson, Roman Ring, Susannah Young, Eliza Rutherford, Tom Hennigan, Jacob Menick, Albin Cassirer, Richard Powell, George van den Driessche, Lisa Anne Hendricks, Maribeth Rauh, Po-Sen Huang, Amelia Glaese, Johannes Welbl, Sumanth Dathathri, Saffron Huang, Jonathan Uesato, John Mellor, Irina Higgins, Antonia Creswell, Nat McAleese, Amy Wu, Erich Elsen, Siddhant Jayakumar, Elena Buchatskaya, David Budden, Esme

Sutherland, Karen Simonyan, Michela Paganini, Laurent Sifre, Lena Martens, Xiang Lorraine Li, Adhiguna Kuncoro, Aida Nematzadeh, Elena Gribovskaya, Domenic Donato, Angeliki Lazaridou, Arthur Mensch, Jean-Baptiste Lespiau, Maria Tsimpoukelli, Nikolai Grigorev, Doug Fritz, Thibault Sottiaux, Mantas Pajarskas, Toby Pohlen, Zhitao Gong, Daniel Toyama, Cyprien de Masson d'Autume, Yujia Li, Tayfun Terzi, Vladimir Mikulik, Igor Babuschkin, Aidan Clark, Diego de Las Casas, Aurelia Guy, Chris Jones, James Bradbury, Matthew Johnson, Blake Hechtman, Laura Weidinger, Iason Gabriel, William Isaac, Ed Lockhart, Simon Osindero, Laura Rimell, Chris Dyer, Oriol Vinyals, Kareem Ayoub, Jeff Stanway, Lorrayne Bennett, Demis Hassabis, Koray Kavukcuoglu, and Geoffrey Irving (2021) Scaling Language Models: Methods, Analysis & Insights from Training Gopher. *Google DeepMind,* https://storage.googleapis.com/deepmind-media/research/language-research/Training%20Goph

Kazi Rafat, Sadia Islam, Abdullah Al Mahfug, Md. Ismail Hossain, Fuad Rahman, Sifat Momen, Shafin Rahman, and Nabeel Mohammed (2023) Mitigating carbon footprint for knowledge distillation based deep learning model compression. *PLoS ONE,* https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0285668

RAIL (n.d.) Responsible AI Licenses. *RAIL,* https://www.licenses.ai/ai-licenses

Inioluwa Deborah Raji, Emily M. Bender, Amandalynne Paullada, Emily Denton, and Alex Hanna (2021) AI and the Everything in the Whole Wide World Benchmark. *35th Conference on Neural Information Processing Systems (NeurIPS 2021). Track on Datasets and Benchmarks,* https://datasets-benchmarks-proceedings.neurips.cc/paper/2021/file/084b6fbb10729ed4da8c3d3f5a3ae7c9-Paper-round2.pdf

Inioluwa Deborah Raji, I. Elizabeth Kumar, Aaron Horowitz, and Andrew Selbst (2022) The Fallacy of AI Functionality. *ACM Digital Library,* https://dl.acm.org/doi/10.1145/3531146.3533158

Inioluwa Deborah Raji, Andrew Smart, Rebecca N. White, Margaret Mitchell, Timnit Gebru, Ben Hutchinson, Jamila Smith-Loud, Daniel Theron, and Parker Barnes (2020) Closing the AI accountability gap: defining an end-to-end framework for internal algorithmic auditing. *In Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (FAT '20). Association for Computing Machinery*, https://dl.acm.org/doi/10.1145/3351095.3372873

Inioluwa Deborah Raji, Peggy Xu, Colleen Honigsberg, and Daniel Ho (2022) Outsider Oversight: Designing a Third Party Audit Ecosystem for AI Governance. *ACM Digital Library,* https://dl.acm.org/doi/10.1145/3514094.3534181

Deepika Raman, Nada Madkour, Evan R. Murphy, Krystal Jackson, and Jessica Newman (2025) Intolerable Risk Threshold Recommendations for Artificial Intelligence. *arXiv*, https://arxiv.org/abs/2503.05812

Alex Ray, Joshua Achiam, and Dario Amodei (2019) Benchmarking Safe Exploration in Deep Reinforcement Learning. *OpenAI,* https://cdn.openai.com/safexp-short.pdf

Tilman Räuker, Anson Ho, Stephen Casper, and Dylan Hadfield-Menell (2023) Toward Transparent AI: A Survey on Interpreting the Inner Structures of Deep Neural Networks. *arXiv*, https://arxiv.org/abs/2207.13243

Scott Reed, Konrad Żołna, Emilio Parisotto, Sergio Gómez Colmenarejo, Alexander Novikov, Gabriel Barth-Maron, Mai Giménez, Yury Sulsky, Jackie Kay, Jost Tobias Springenberg, Tom Eccles, Jake Bruce, Ali Razavi, Ashley Edwards, Nicolas Heess, Yutian Chen, Raia Hadsell, Oriol Vinyals, Mahyar Bordbar, and Nando de Freitas (2022) A Generalist Agent. *Google DeepMind,* https://storage.googleapis.com/deepmind-media/A%20Generalist%20Agent/Generalist%20Agent.pdf

David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R. Bowman (2023) GPQA: A Graduate-Level Google-Proof Q&A Benchmark. *arXiv*, https://arxiv.org/abs/2311.12022

Richard Ren, Steven Basart, Adam Khoja, Alexander Pan, Alice Gatti, Long Phan, Xuwang Yin, Mantas Mazeika, Gabriel Mukobi, Ryan Hwang Kim, Stephen Fitz, and Dan Hendrycks (2024) Safetywashing: Do AI Safety Benchmarks Actually Measure Safety Progress? *OpenReview,* https://openreview.net/pdf?id=YagfTP3RK6

Anka Reuel, Amelia Hardy, Chandler Smith, Max Lamparth, Malcolm Hardy, and Mykel J. Kochenderfer (2024) BetterBench: Assessing AI Benchmarks, Uncovering Issues, and Establishing Best Practices. *arXiv*, https://arxiv.org/abs/2411.12990

Marina Ribeiro, Bárbara Malcorra, Natália B. Mota, Rodrigo Wilkens, Aline Villavicencio, Lilian C. Hubner, and César Rennó-Costa (2024) A Methodology for Explainable Large Language Models with Integrated Gradients and Linguistic Analysis in Text Classification. *arXiv,* https://arxiv.org/pdf/2410.00250

Dan Ristea and Vasilios Mavroudis (2025) HonestCyberEval: An AI Cyber Risk Benchmark for Automated Software Exploitation. *arXiv,* https://arxiv.org/abs/2410.21939

Alexander Robey, Zachary Ravichandran, Vijay Kumar, Hamed Hassani, and George J. Pappas (2025) Jailbreaking LLM-Controlled Robots. *International Conference on Robotics and Automation (ICRA) 2025,* https://robopair.org/

Ben Robinson and James Ginns (2024) Transforming risk governance at frontier AI companies. *The Centre for Long-Term Resilience,* https://www.longtermresilience.org/wp-content/uploads/2024/07/Transforming-risk-governance-at-frontier-AI-companies-CLTR-1.pdf

Fabien Roger, Ryan Greenblatt, Max Nadeau, Buck Shlegeris, and Nate Thomas (2023) Benchmarks for Detecting Measurement Tampering. *arXiv*, https://arxiv.org/abs/2308.15605

Sayak Saha Roy, Poojitha Thota, Krishna Vamsi Naragam,and Shirin Nilizadeh (2024) From Chatbots to PhishBots? - Preventing Phishing scams created using ChatGPT, Google Bard and Claude. *arXiv,* https://arxiv.org/pdf/2310.19181

Yangjun Ruan, Chris J. Maddison, and Tatsunori Hashimoto (2024) Observational Scaling Laws and the Predictability of Language Model Performance. *arXiv,* https://arxiv.org/abs/2405.10938

Rachel Rudinger, Jason Naradowsky, Brian Leonard, and Benjamin Van Durme (2019) Winogender Schemas. *GitHub*, https://github.com/rudinger/winogender-schemas

Tim G. J. Rudner and Helen Toner (2021) Key Concepts in AI Safety: Specification in Machine Learning. *Georgetown University Center for Security and Emerging Technology,* https://cset.georgetown.edu/wp-content/uploads/Key-Concepts-in-AI-Safety-Specification-in-Machine-Learning.pdf

Stuart Russell (2019) *Human Compatible: Artificial Intelligence and the Problem of Control*. Viking

Malak Sadek, Marios Constantinides, Daniele Quercia, and Celine Mougenot (2024) Guidelines for integrating value sensitive design in responsible AI toolkits. *arXiv*, https://arxiv.org/pdf/2403.00145v1

Alexander K. Saeri, Sophia Lloyd George, Jess Graham, Clelia Lacarriere, Peter Slattery, and Neil Thompson (2025) Mapping AI Risk Mitigations: Evidence Scan and Draft Mitigation Taxonomy. *MIT AI Risk Initiative,* https://airisk.mit.edu/blog/mapping-ai-risk-mitigations

SaferAI (2025) Informing the Code of Practice: A Hierarchical Methodology of Defining Risk Tiers. *SaferAI,* https://www.safer-ai.org/informing-the-code-of-practice-a-hierarchical-methodology-of-defining-risk-tiers

Oscar Sainz, Jon Campos, Iker García-Ferrero, Julen Etxaniz, Oier Lopez de Lacalle, Eneko Agirre (2023) NLP Evaluation in trouble: On the Need to Measure LLM Data Contamination for Each Benchmark. *arXiv*, https://aclanthology.org/2023.findings-emnlp.722/

Nader Salari, Mahan Beiromvand, Amin Hosseinian-Far, Javad Habibi, Fateme Babajani, and Masoud Mohammadi (2025) Impacts of Generative Artificial Intelligence on The Future of Labor Market: A Systematic Review. *ScienceDirect,* https://www.sciencedirect.com/science/article/pii/S2451958825000673

Pedro Saleiro, Benedict Kuester, Abby Stevens, Ari Anisfeld, Loren Hinkson, Jesse London, and Rayid Ghani (2019) Aequitas: A Bias and Fairness Audit Toolkit. *arXiv,* https://arxiv.org/abs/1811.05577

Pamela Samuelson (2023) Generative AI meets copyright: Ongoing lawsuits could affect everyone who uses generative AI. *Science 381 (6654) 158-161,* https://www.science.org/doi/10.1126/science.adi0656

Jonas B. Sandbrink (2023) Artificial intelligence and biological misuse: Differentiating risks of language models and biological design tools. *arXiv*, https://arxiv.org/abs/2306.13952

Tony Berber Sardinha (2024) AI-generated vs Human-Authored Texts: A Multidimensional Comparison. *ScienceDirect,* https://doi.org/10.1016/j.acorp.2023.100083

Srinivasan Sastry, Lukas Heim, Henry Belfield, Max Anderljung, Miles Brundage, James Hazell, Chris O'Keefe, Geoffrey K. Hadfield, Ryan Ngo, Kieran Pilz, Giorgos Gor, Emma Bluemke, Sam Shoker, Jack Egan, Robert F. Trager, Sacha Avin, Anna Weller, Yoshua Bengio, and Daniel Coyle (2024) Computing power and the governance of artificial intelligence. *arXiv*, https://arxiv.org/abs/2402.08797

Rylan Schaeffer, Brando Miranda, and Sanmi Koyejo (2023) Are Emergent Abilities of Large Language Models a Mirage? *arXiv*, https://arxiv.org/abs/2304.15004

Jeremy Scheurer, Mikita Balesni, and Marius Hobbhahn (2024) Large Language Models can Strategically Deceive their Users when Put Under Pressure. *arXiv*, https://arxiv.org/abs/2311.07590

Victor Schmidt, Alexandra (Sasha) Luccioni, Alexandre Lacoste, and Thomas Dandres (2019) ML CO2 Impact. *GitHub*, https://mlco2.github.io/impact/

Marc Schmitt and Ivan Flechais (2024) Digital Deception: Generative Artificial Intelligence in Social Engineering and Phishing. *SpringerNature,* https://link.springer.com/article/10.1007/s10462-024-10973-2

Bronson Schoen, Evgenia Nitishinskaya, Mikita Balesni, Axel Højmark, Felix Hofstätter, Jérémy Scheurer, Alexander Meinke, Jason Wolfe, Teun van der Weij, Alex Lloyd, Nicholas Goldowsky-Dill, Angela Fan, Andrei Matveiakin, Rusheb Shah, Marcus Williams, Amelia Glaese, Boaz Barak, Wojciech Zaremba, and Marius Hobbhahn (2025) Stress Testing Deliberative Alignment for Anti-Scheming Training. *arXiv*, https://www.arxiv.org/abs/2509.15541

Jonas Schuett (2022) Three lines of defense against risks from AI. *arXiv*, https://arxiv.org/abs/2212.08364

Jonas Schuett (2023) Three lines of defense against risks from AI. *Springer Nature*, https://link.springer.com/article/10.1007/s00146-023-01811-0

Jonas Schuett, Noemi Dreksler, Markus Anderljung, David McCaffary, Lennart Heim, Emma Bluemke, and Ben Garfinkel (2023) Towards best practices in AGI safety and governance: A survey of expert opinion. *arXiv*, https://arxiv.org/abs/2305.07153

Reva Schwartz, Apostol Vassilev, Kristen Greene, Lori Perine, Andrew Burt, and Patrick Hall (2022) Towards a Standard for Identifying and Managing Bias in Artificial Intelligence. SP 1270. *National Institute of Standards and Technology,* https://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.1270.pdf

Reva Schwartz, Jonathan Fiscus, Kristen Greene, Gabriella Waters, Rumman Chowdhury, Theodore Jensen, Craig Greenberg, Afzal Godil, Razvan Amironesei, Patrick Hall, and Shomik Jain (2024) The NIST Assessing Risks and Impacts of AI (ARIA) Pilot Evaluation Plan. *National Institute of Standards and Technology,* https://ai-challenges.nist.gov/uassets/7

Elizabeth Seger, Noemi Dreksler, Richard Moulange, Emily Dardaman, Jonas Schuett, K. Wei, Christoph Winter, Mackenzie Arnold, Seán Ó hÉigeartaigh, Anton Korinek, Markus Anderljung, Ben Bucknall, Alan Chan, Eoghan Stafford, Leonie Koessler, Aviv Ovadya, Ben Garfinkel, Emma Bluemke, Michael Aird, Patrick Levermore, Julian Hazell, and Abhishek Gupta (2023) Open-Sourcing Highly Capable Foundation Models: An Evaluation of Risks, Benefits, and Alternative Methods for Pursuing Open-Source Objectives. *Centre for the Governance of AI,* https://cdn.governance.ai/Open-Sourcing_Highly_Capable_Foundation_Models_2023_GovAI.pdf

Rohin Shah, Vikrant Varma, Ramana Kumar, Mary Phuong, Victoria Krakovna, Jonathan Uesato, and Zac Kenton (2022) Goal Misgeneralization: Why Correct Specifications Aren't Enough For Correct Goals. *arXiv*, https://arxiv.org/abs/2210.01790

Minghao Shao, Sofija Jancheska, Meet Udeshi, Brendan Dolan-Gavitt, Haoran Xi, Kimberkly Milner, Boyuan Chen, Max Yin, Siddharth Garg, Prashanth Krishnamurthy, Parshad Korrami, Ramesh Karri, and Muhammad Shafique (2025) NYU CTF Bench: A Scalable Open-Source Benchmark Dataset for Evaluating LLMs in Offensive Security. *arXiv,* https://arxiv.org/pdf/2406.05590

Lee Sharkey, Clíodhna Ghuidhir, Dan Braun, Jérémy Scheurer, Mikita Balesni, Lucius Bushnaq, Charlotte Stix, and Marius Hobbhahn (2024) A Causal Framework for AI Regulation and Auditing. *Preprints,* https://www.preprints.org/manuscript/202401.1424/v1

Matthew Sharp, Omer Bilgin, Iason Gabriel, and Lewis Hammond (2025) Agentic Inequality. *arXiv*, https://www.arxiv.org/abs/2510.16853

Hong Shen, Alicia DeVos, Motahhare Eslami, and Kenneth Holstein (2021) Everyday Algorithm Auditing: Understanding the Power of Everyday Users in Surfacing Harmful Algorithmic Behaviors. *ACM Digital Library,* https://dl.acm.org/doi/10.1145/3479577

Toby Shevlane (2022) Structured access: an emerging paradigm for safe AI deployment. *arXiv*, https://arxiv.org/abs/2201.05159

Toby Shevlane, Sebastian Farquhar, Ben Garfinkel, Mary Phuong, Jess Whittlestone, Jade Leung, Daniel Kokotajlo, Nahema Marchal, Markus Anderljung, Noam Kolt, Lewis Ho, Divya Siddarth, Shahar Avin, Will Hawkins, Been Kim, Iason Gabriel, Vijay Bolina, Jack Clark, Yoshua Bengio, Paul Christiano, and Allan Dafoe (2023) Model evaluation for extreme risks. *arXiv,* https://arxiv.org/abs/2305.15324

Eran Shimony and Omer Tsarfati (2023) Chatting Our Way Into Creating a Polymorphic Malware. *CyberArk,* https://www.cyberark.com/resources/threat-research-blog/chatting-our-way-into-creating-a-polymorphic-malware

Noah Shinn, Beck Labash, and Ashwin Gopinath (2023). Reflexion: an autonomous agent with dynamic memory and self-reflection. *arXiv*, https://arxiv.org/abs/2303.11366

Ilia Shumailov, Zakhar Shumaylov, Yiren Zhao, Yarin Gal, Nicolas Papernot, and Ross Anderson (2024) The Curse of Recursion: Training Generated Data Makes Models Forget. *arXiv*, https://arxiv.org/abs/2305.17493

Significant Gravitas (2023) Auto-GPT. *GitHub*, https://github.com/Significant-Gravitas/Auto-GPT

Barnaby Simkin, Nikki Pope, Leon Derczynski, and Christopher Parisien (2025) Frontier AI Risk Assessment. *Nvidia,* https://images.nvidia.com/content/pdf/NVIDIA-Frontier-AI-Risk-Assessment.pdf

Ranjit Singh, Borhane Blili-Hamelin, Carol Anderson, Emnet Tafesse, Briana Vecchione, Beth Duckles, and Jacob Metcalf (2025) Red-Teaming in the Public Interest. *Data & Society*, https://ranjitsingh.me/wp-content/uploads/2025/02/Singh_Red-Teaming-in_the_Public_Interest.pdf

Peter Slattery, Alexander K. Saeri, Emily A. C. Grundy, Jess Graham, Michael Noetel, Risto Uuk, James Dao, Soroush Pour, Stephen Casper, and Neil Thompson (2025) The AI Risk Repository: A Comprehensive Meta-Review, Database, and Taxonomy of Risks From Artificial Intelligence. *arXiv*, https://arxiv.org/pdf/2408.12622

SLSA (n.d.) Safeguarding Artifact Integrity Across any Software Supply Chain. *SLSA,* https://slsa.dev/

Emily H. Soice, Rafael Rocha, Kimberlee Cordova, Michael Specter, and Kevin M. Esvelt (2023) Can large language models democratize access to dual-use biotechnology? *arXiv*, https://arxiv.org/abs/2306.03809

Irene Solaiman, Miles Brundage, Jack Clark, Amanda Askell, Ariel Herbert-Voss, Jeff Wu, Alec Radford, Gretchen Krueger, Jong Wook Kim, Sarah Kreps, Miles McCain, Alex Newhouse, Jason Blazakis, Kris McGuffie, and Jasmine Wang (2019) Release Strategies and the Social Impacts of Language Models. *arXiv*, https://arxiv.org/abs/1908.09203

Irene Solaiman (2023) The Gradient of Generative AI Release: Methods and Considerations. *arXiv,* https://arxiv.org/abs/2302.04844

Irene Solaiman, Zeerak Talat, William Agnew, Lama Ahmad, Dylan Baker, Su Lin Blodgett, Hal Daumé III, Jesse Dodge, Ellie Evans, Sara Hooker, Yacine Jernite, Alexandra Sasha Luccioni, Alberto Lusoli, Margaret Mitchell, Jessica Newman, Marie-Therese Png, Andrew Strait, and Apostol Vassilev (2023) Evaluating the Social Impact of Generative AI Systems in Systems and Society. *arXiv*, https://arxiv.org/abs/2306.05949

Murugiah Souppaya (NIST), Karen Scarfone (Scarfone Cybersecurity), and Donna Dodson (2022) Secure Software Development Framework (SSDF) Version 1.1: Recommendations for Mitigating the Risk of Software Vulnerabilities. NIST SP 800-218. *National Institute of Standards and Technology*, https://csrc.nist.gov/pubs/sp/800/218/final

SPDX (n.d.a) SPDX AI. *SPDX,* https://spdx.dev/learn/areas-of-interest/ai/

SPDX (n.d.b) The System Package Data Exchange™ (SPDX®). *SPDX,* https://spdx.dev/

Madhulika Srikumar, Jiyoo Chang, and Kasia Chmielinski (2024) Risk Mitigation Strategies for the Open Foundation Model Value Chain. *Partnership on AI,* https://partnershiponai.org/resource/risk-mitigation-strategies-for-the-open-foundation-model-value-chain/

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R. Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, Agnieszka Kluska, Aitor Lewkowycz, Akshat Agarwal, Alethea Power, Alex Ray, Alex Warstadt, Alexander W. Kocurek, Ali Safaya, Ali Tazarv, Alice Xiang, Alicia Parrish, Allen Nie, Aman Hussain, Amanda Askell, Amanda Dsouza, Ambrose Slone, Ameet Rahane, Anantharaman S. Iyer, Anders Andreassen, Andrea Madotto, Andrea Santilli, Andreas Stuhlmüller, Andrew Dai, Andrew La, Andrew Lampinen, Andy Zou, Angela Jiang, Angelica Chen, Anh Vuong, Animesh Gupta, Anna Gottardi, Antonio Norelli, Anu Venkatesh, Arash Gholamidavoodi, Arfa Tabassum, Arul Menezes, Arun Kirubarajan, Asher Mullokandov, Ashish Sabharwal, Austin Herrick, Avia Efrat, Aykut Erdem, Ayla Karakaş, B. Ryan Roberts, Bao Sheng Loe, Barret Zoph, Bartłomiej Bojanowski, Batuhan Özyurt, Behnam Hedayatnia, Behnam Neyshabur, Benjamin Inden, Benno Stein, Berk Ekmekci, Bill Yuchen Lin, Blake Howald, Cameron Diao, Cameron Dour, Catherine Stinson, Cedrick Argueta, César Ferri Ramírez, Chandan Singh, Charles Rathkopf, Chenlin Meng, Chitta Baral, Chiyu Wu, Chris Callison-Burch, Chris Waites, Christian Voigt, Christopher D. Manning, Christopher Potts, Cindy Ramirez, Clara E. Rivera, Clemencia Siro, Colin Raffel, Courtney Ashcraft, Cristina Garbacea, Damien Sileo, Dan Garrette, Dan Hendrycks, Dan Kilman, Dan Roth, Daniel Freeman, Daniel Khashabi, Daniel Levy, Daniel Moseguí González, Danielle Perszyk, Danny Hernandez, Danqi Chen, Daphne Ippolito, Dar Gilboa et al. (2022) Beyond the Imitation Game: Quantifying and extrapolating the capabilities of language models. *arXiv*, https://arxiv.org/abs/2206.04615

Merlin Stein and Connor Dunlop (2024) Safe Beyond Sale: Post-Deployment Monitoring of AI. *Ada Lovelace Institute,* https://www.adalovelaceinstitute.org/blog/post-deployment-monitoring-of-ai/

Jacob Steinhardt and Beth Barnes (2021) Measuring and forecasting risks. https://docs.google.com/document/d/1cPwcUSloY8TyZxCumGPBhdVUNoYyyw9AR1QshlRI3gc/edit?usp=sharing

Kevin Stine, Stephen Quinn, Gregory Witte, and Robert Gardner (2020) Integrating Cybersecurity and Enterprise risk-management (ERM). NISTIR 8286. *National Institute of Standards and Technology,* https://csrc.nist.gov/publications/detail/nistir/8286/final

Laura Stroebel, Mark Llewellyn, Tricia Hartley, Tsui Shan Ip, and Mohiuddin Ahmed (2023) A Systematic Literature review on the Effectiveness of Deepfake Detection Techniques. *Journal of Internet Commerce,* https://www.tandfonline.com/doi/full/10.1080/23742917.2023.2192888#abstract

Hsuan Su, Cheng-Chu Cheng, Hua Farn, Shachi H Kumar, Saurav Sahay, Shang-Tse Chen, and Hung-yi Lee (2023) Learning from Red Teaming: Gender Bias Provocation and Mitigation in Large Language Models. *arXiv,* https://arxiv.org/abs/2310.11079v1

TAIBOM (n.d.) Trustable AI Bill of Materials. *NquiringMinds,* https://taibom.org/

Rishub Tamirisa, Bhrugu Bharathi, Long Phan, Andy Zhou, Alice Gatti, Tarun Suresh, Maxwell Lin, Justin Wang, Rowan Wang, Ron Arel, Andy Zou, Dawn Song, Bo Li, Dan Hendrycks, and Mantas Mazeika (2025) Tamper-Resistant Safeguards for Open-Weight LLMs. *arXiv,* https://arxiv.org/pdf/2408.00761

Bill: HB 149 (2025) *Texas Legislature,* https://capitol.texas.gov/BillLookup/History.aspx?LegSess=89R&Bill=HB149

TFS (2025) Serious Incident Prevention for AI: Lessons From Other Industries and Recommendations for the EU AI Office. *The Future Society,* https://thefuturesociety.org/serious-ai-incident-prevention-eu

The Associated Press (2023) Elon Musk says he'll create 'TruthGPT' to counter AI 'bias'. The *Independent,* https://www.independent.co.uk/news/world/americas/elon-musk-ap-chatgpt-tesla-openai-b2321564.html

Helen Toner (2023) What Are Generative AI, Large Language Models, and Foundation Models? *Georgetown University Center for Security and Emerging Technology,* https://cset.georgetown.edu/article/what-are-generative-ai-large-language-models-and-foundation-models/

Violet Turri and Rachel Dzombak (2023) Why We Need to Know More: Exploring the State of AI Incident Documentation Practices. *In AAAI/ACM Conference on AI, Ethics, and Society (AIES '23). ACM Digital Library,* https://dl.acm.org/doi/10.1145/3600211.3604700

Laura D. Tyson and John Zysman (2022) Automation, AI & Work. *Daedalus* 151 (2): 256–271. *MIT Press Direct*, https://doi.org/10.1162/daed_a_01914

UK AISI (n.d.) Inspect: An Open-Source Framework for Large Language Model Evaluations. *UK Artificial Intelligence Security Institute,* https://inspect.aisi.org.uk/

UN (1948) Universal Declaration of Human Rights (UDHR). *United Nations*, https://www.un.org/en/about-us/universal-declaration-of-human-rights

UN (2011) UN Guiding Principles on Business and Human Rights: Implementing the United Nations "Protect, Respect and Remedy" Framework. *United Nations Office of the High Commissioner on Human Rights,* https://www.ohchr.org/documents/publications/guidingprinciplesbusinesshr_en.pdf

UNESCO (2023) Ethical impact assessment: a tool of the Recommendation on the Ethics of Artificial Intelligence. *United Nations Educational, Scientific and Cultural Organization,* https://doi.org/10.54678/YTSA7796

Karthik Valmeekam, Matthew Marquez, Alberto Olmo, Sarath Sreedharan, Subbarao Kambhampati (2022a) PlanBench: An Extensible Benchmark for Evaluating Large Language Models on Planning and Reasoning about Change. *arXiv*, https://arxiv.org/abs/2206.10498

Karthik Valmeekam, Matthew Marquez, Alberto Olmo, Sarath Sreedharan, and Subbarao Kambhampati (2022b) PlanBench: An Extensible Benchmark for Evaluating Large Language Models on Planning and Reasoning about Change. *GitHub*, https://github.com/karthikv792/LLMs-Planning/tree/main/plan-bench

Apostol Vassilev, Alina Oprea, Alie Fordyce, Hyrum Anderson, Xander Davies, and Maia Hamin (2025) NIST AI 100-2e2025, Adversarial Machine Learning: A Taxonomy and Terminology of Attacks and Mitigations. *National Institute of Standards and Technology,* https://csrc.nist.gov/pubs/ai/100/2/e2025/final

David Waltz and Bruce G Buchanan (2009) Automating Science. *Science, 324(5923), 43–44,* https://doi.org/10.1126/science.1172781

Yixin Wan and Kai-Wei Chang (2025) White Men Lead, Black Women Help? Benchmarking and Mitigating Language Agency Social Biases in LLMs. *arXiv*, https://arxiv.org/abs/2404.10508

Shengye Wan, Cyrus Nikolaidis, Daniel Song, David Molnar, James Crnkovich, Jayson Grace, Manish Bhatt, Sahana Chennabasappa, Spencer Whitman, Stephanie Ding, Vlad Ionescu, Yue Li, and Joshua Saxe (2024) CYBERSECEVAL 3: Advancing the Evaluation of Cybersecurity Risks and Capabilities in Large Language Models. *arXiv*, https://arxiv.org/abs/2408.01605

Boxin Wang, Weixin Chen, Hengzhi Pei, Chulin Xie, Mintong Kang, Chenhui Zhang, Chejian Xu, Zidi Xiong, Ritik Dutta, Rylan Schaeffer, Sang T. Truong, Simran Arora, Mantas Mazeika, Dan Hendrycks, Zinan Lin, Yu Cheng, Sanmi Koyejo, Dawn Song, and Bo Li (2023) DecodingTrust: A Comprehensive Assessment of Trustworthiness in GPT Models. *arXiv*, https://arxiv.org/abs//2306.11698

Yuxia Wang, Haonan Li, Xudong Han, Preslav Nakov, and Timothy Baldwin (2023) Do-Not-Answer: A Dataset for Evaluating Safeguards in LLMs. *arXiv*, https://arxiv.org/abs/2308.13387

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman (2019) SuperGLUE: A Stickier Benchmark for General-Purpose Language Understanding Systems. *In 33rd Conference on Neural Information Processing Systems (NeurIPS 2019)*, https://w4ngatang.github.io/static/papers/superglue.pdf

Zeyu Wang (2024) CausalBench: A Comprehensive Benchmark for Evaluating Causal Reasoning Capabilities of Large Language Models. *Proceedings of the 10th SIGHAN Workshop on Chinese Language Processing,* https://aclanthology.org/2024.sighan-1.17.pdf

Zhun Wang, Tianneng Shi, Jingxuan He, Matthew Cai, Jialin Zhang, and Dawn Song (2025) CyberGym: Evaluating AI Agents' Real-World Cybersecurity Capabilities at Scale. *arXiv*, https://arxiv.org/abs/2506.02548

Hilde Weerts, Florian Pfisterer, Matthias Feurer, Katharina Eggensperger, Edward Bergman, Noor Awad, Joaquin Vanschoren, Mykola Pechenizkiy, Bernd Bischl, and Frank Hutter (2024) Can Fairness be Automated? Guidelines and Opportunities for Fairness-Aware AutoML. *Journal of Artificial Intelligence Research,* https://jair.org/index.php/jair/article/view/14747

WEF (2024) Navigating the AI Frontier: A Primer on the Evolution and Impact of AI Agents. *World Economic Forum,* https://reports.weforum.org/docs/WEF_Navigating_the_AI_Frontier_2024.pdf

WEF (2025a) AI Agents in Action: Foundations for Evaluation and Governance. *World Economic Forum,* https://www.weforum.org/publications/ai-agents-in-action-foundations-for-evaluation-and-governance/

WEF (2025b) Global Risks Report 2025. *World Economic Forum,* https://www.weforum.org/publications/global-risks-report-2025/

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus (2022) Emergent Abilities of Large Language Models. *Transactions on Machine Learning Research,* https://openreview.net/pdf?id=yzkSU5zdwD

Laura Weidinger, Joslyn Barnhart, Jenny Brennan, Christina Butterfield, Susie Yong, Will Hawkins, Lisa Anne Hendricks, Ramona Comanescu, Oscar Chang, Mikel Rodriguez, Jennifer Beroshi, Dawn Bloxwich, Lev Proleev, Jilin Chen, Sebastian Farquhar, Lewis Ho, Iason Gabriel, Allan Dafoe, and William Isaac (2024) Holistic Safety and Responsibility Evaluations of Advanced AI Models. *arXiv,* https://arxiv.org/abs/2404.14068

Laura Weidinger, John Mellor, Maribeth Rauh, Conor Griffin, Jonathan Uesato, Po-Sen Huang, Myra Cheng, Mia Glaese, Borja Balle, Atoosa Kasirzadeh, Zac Kenton, Sasha Brown, Will Hawkins, Tom Stepleton, Courtney Biles, Abeba Birhane, Julia Haas, Laura Rimell, Lisa Anne Hendricks, William Isaac, Sean Legassick, Geoffrey Irving, and Iason Gabriel (2021) Ethical and Social Risks of Harm from Language Models. *arXiv*, https://arxiv.org/abs/2112.04359

Laura Weidinger, Maribeth Rauh, Nahema Marchal, Arianna Manzini, Lisa Anne Hendricks, Juan Mateos-Garcia, Stevie Bergman, Jackie Kay, Conor Griffin, Ben Bariach, Iason Gabriel, Verena Rieser, and William Isaac (2023) Sociotechnical Safety Evaluation of Generative AI Systems. *arXiv*, https://arxiv.org/abs/2310.11986

Laura Weidinger, Jonathan Uesato, Maribeth Rauh, Conor Griffin, Po-Sen Huang, John Mellor, Amelia Glaese, Myra Cheng, Borja Balle, Atoosa Kasirzadeh, Courtney Biles, Sasha Brown, Zac Kenton, Will Hawkins, Tom Stepleton, Abeba Birhane, Lisa Anne Hendricks, Laura Rimell, William Isaac, Julia Haas, Sean Legassick, Geoffrey Irving, and Iason Gabriel (2022) Taxonomy of Risks posed by Language Models. *In 2022 ACM Conference on Fairness, Accountability, and Transparency* (*FAccT '22*). *Association for Computing Machinery, New York, NY, USA, 214–229,* https://doi.org/10.1145/3531146.3533088

Laura Weidinger, John Mellor, Bernat Guillen Pegueroles, Nahema Marchal, Ravin Kumar, Kristian Lum, Canfer Akbulut, Mark Diaz, Stevie Bergman, Mikel Rodriguez, Verena Rieser, and William Isaac (2024) STAR: SocioTechnical Approach to Red Teaming Language Models. *arXiv*, https://arxiv.org/abs/2406.11757

Teun van der Weij, Felix Hofstätter, Ollie Jaffe, Samuel F. Brown, and Francis Rhys Ward (2024) AI Sandbagging: Language Models can Strategically Underperform on Evaluations. *arXiv*, https://arxiv.org/abs/2406.07358

White House (2023a) Ensuring Safe, Secure, and Trustworthy AI. *White House,* https://bidenwhitehouse.archives.gov/wp-content/uploads/2023/07/Ensuring-Safe-Secure-and-Trustworthy-AI.pdf

Hjalmar Wijk, Tao Lin, Joel Becker, Sami Jawhar, Neev Parikh, Thomas Broadley, Lawrence Chan, Michael Chen, Josh Clymer, Jai Dhyani, Elena Ericheva, Katharyn Garcia, Brian Goodrich, Nikola Jurkovic, Holden Karnofsky, Megan Kinniment, Aron Lajko, Seraphina Nix, Lucas Sato, William Saunders, Maksym Taran, Ben West, and Elizabeth Barnes (2025) RE-Bench: Evaluating Frontier AI R&D Capabilities of Language Model Agents Against Human Experts. *arXiv*, https://arxiv.org/abs/2411.15114

Sophie Williams, Noemi Dreksler, Aidan Homewood, Markus Anderljung, and Jonas Freund (2025) Assessing Risk Relative to Competitors: An Analysis of Current AI Company Policies. *Centre for the Governance of AI,* https://cdn.governance.ai/Assessing_Risk_Relative_to_Competitors_An_Analysis_of_Current_AI_Company_Policies.pdf

Amy Winecoff and Miranda Bogen (2024) Best Practices in AI Documentation: The Imperative of Evidence from Practice. *CDT AI Governance Lab,* https://cdt.org/insights/best-practices-in-ai-documentation-the-imperative-of-evidence-from-practice/

Anna Katariina Wisakanto, Joe Rogero, Avyay M. Casheekar, and Richard Mallah (2025) Adapting Probabilistic Risk Assessment for AI. *arXiv*, https://arxiv.org/pdf/2504.18536

Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Beibin Li, Erkang Zhu, Li Jiang, Xiaoyun Zhang, Shaokun Zhang, Jiale Liu, Ahmed Hassan Awadallah, Ryen W White, Doug Burger, and Chi Wang (2023) AutoGen: Enabling Next-Gen LLM Applications via Multi-Agent Conversation. *arXiv,* https://arxiv.org/abs/2308.08155

xAI (2024) Open Release of Grok-1. *xAI,* https://x.ai/blog/grok-os

Zhiheng Xi, Wenxiang Chen, Xin Guo, Wei He, Yiwen Ding, Boyang Hong, Ming Zhang, Junzhe Wang, Senjie Jin, Enyu Zhou, Rui Zheng, Xiaoran Fan, Xiao Wang, Limao Xiong, Yuhao Zhou, Weiran Wang, Changhao Jiang, Yicheng Zou, Xiangyang Liu, Zhangyue Yin, Shihan Dou, Rongxiang Weng, Wensen Cheng, Qi Zhang, Wenjuan Qin, Yongyan Zheng, Xipeng Qiu, Xuanjing Huang, and Tao Gui (2023) The Rise and Potential of Large Language Model Based Agents: A Survey. *arXiv,* https://arxiv.org/abs/2309.07864

Tinghao Xie, Xiangyu Qi, Yi Zeng, Yangsibo Huang, Udari Madhushani Sehwag, Kaixuan Huang, Luxi He, Boyi Wei, Dacheng Li, Ying Sheng, Ruoxi Jia, Bo Li, Kai Li, Danqi Chen, Peter Henderson, and Prateek Mittal (2025) SORRY-Bench: Systematically Evaluating Large Language Model Safety Refusal. *arXiv*, https://arxiv.org/abs/2406.14598

Ruijie Xu, Zengzhi Wang, Run-Ze Fan, and Pengfei Liu (2024) Benchmarking Benchmark Leakage in Large Language Models. *arXiv*, https://arxiv.org/abs/2404.18824

Shunyu Yao, Noah Shinn, Pedram Razavi, and Karthik Narasimhan (2024) τ-bench: A Benchmark for Tool-Agent-User Interaction in Real-World Domains. *arXiv*, https://arxiv.org/abs/2406.12045

Bill Yuchen Lin, Yuntian Deng, Khyathi Chandu, Faeze Brahman, Abhilasha Ravichander, Valentina Pyatkin, Nouha Dziri, Ronan Le Bras, and Yejin Choi (2024) WildBench: Benchmarking LLMs with Challenging Tasks from Real Users in the Wild. *arXiv*, https://arxiv.org/abs/2406.04770

Marisa Zalabak, Balaji Dhamodharan, Bill Lesieur, Olga Magnusson, Shannon Kennedy, and Sundar Krishnan (2025) The ROI of AI Ethics: Profiting with Principles for the Future. *The Digital Economist,* https://www.aigl.blog/content/files/2025/06/The-ROI-of-AI-Ethics-Profiting-with-Principles-for-the-Future.pdf

Maxwell Zeff (2025) Anthropic's New AI Model Turns to Blackmail When Engineers Try to Take it Offline. *TechCrunch,* https://techcrunch.com/2025/05/22/anthropics-new-ai-model-turns-to-blackmail-when-engineers-try-to-take-it-offline/

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, Todor Mihaylov, Myle Ott, Sam Shleifer, Kurt Shuster, Daniel Simig, Punit Singh Koura, Anjali Sridhar, Tianlu Wang, and Luke Zettlemoyer (2022) OPT: Open Pre-trained Transformer Language Models. *arXiv*, https://arxiv.org/abs/2205.01068

Hugh Zhang, Jeff Da, Dean Lee, Vaughn Robinson, Catherine Wu, Will Song, Tiffany Zhao, Pranav Raja, Charlotte Zhuang, Dylan Slack, Qin Lyu, Sean Hendryx, Russell Kaplan, Michele Lunati, and Summer Yue (2024) A Careful Examination of Large Language Model Performance on Grade School Arithmetic. *arXiv*, https://arxiv.org/abs/2405.00332

Andy K. Zhang, Neil Perry, Riya Dulepet, Joey Ji, Celeste Menders, Justin W. Lin, Eliot Jones, Gashon Hussein, Samantha Liu, Donovan Jasper, Pura Peetathawatchai, Ari Glenn, Vikram Sivashankar, Daniel Zamoshchin, Leo Glikbarg, Derek Askaryar, Mike Yang, Teddy Zhang, Rishi Alluri, Nathan Tran, Rinnara Sangpisit, Polycarpos Yiorkadjis, Kenny Osele, Gautham Raghupathi, Dan Boneh, Daniel E. Ho, and Percy Liang (2024) Cybench: A Framework for Evaluating Cybersecurity Capabilities and Risks of Language Models. *arXiv,* https://arxiv.org/abs/2408.08926

Wanjun Zhong, Ruixiang Cui, Yiduo Guo, Yaobo Liang, Shuai Lu, Yanlin Wang, Amin Saied, Weizhu Chen, and Nan Duan (2023) AGIEval: A Human-Centric Benchmark for Evaluating Foundation Models. *arXiv*, https://arxiv.org/abs/2304.06364

Andy Zou, Zifan Wang, J. Zico Kolter, and Matt Fredrikson (2023a) Universal and Transferable Adversarial Attacks on Aligned Language Models. *arXiv*, https://arxiv.org/abs/2307.15043

Andy Zou, Zifan Wang, J. Zico Kolter, and Matt Fredrikson (2023b) LLM Attacks. *GitHub*, https://github.com/llm-attacks/llm-attacks

Remco Zwetsloot and Allan Dafoe (2019) Thinking About Risks From AI: Accidents, Misuse and Structure. *Lawfare,* https://www.lawfareblog.com/thinking-about-risks-ai-accidents-misuse-and-structure

CLTC

Center for Long-Term Cybersecurity

UC Berkeley