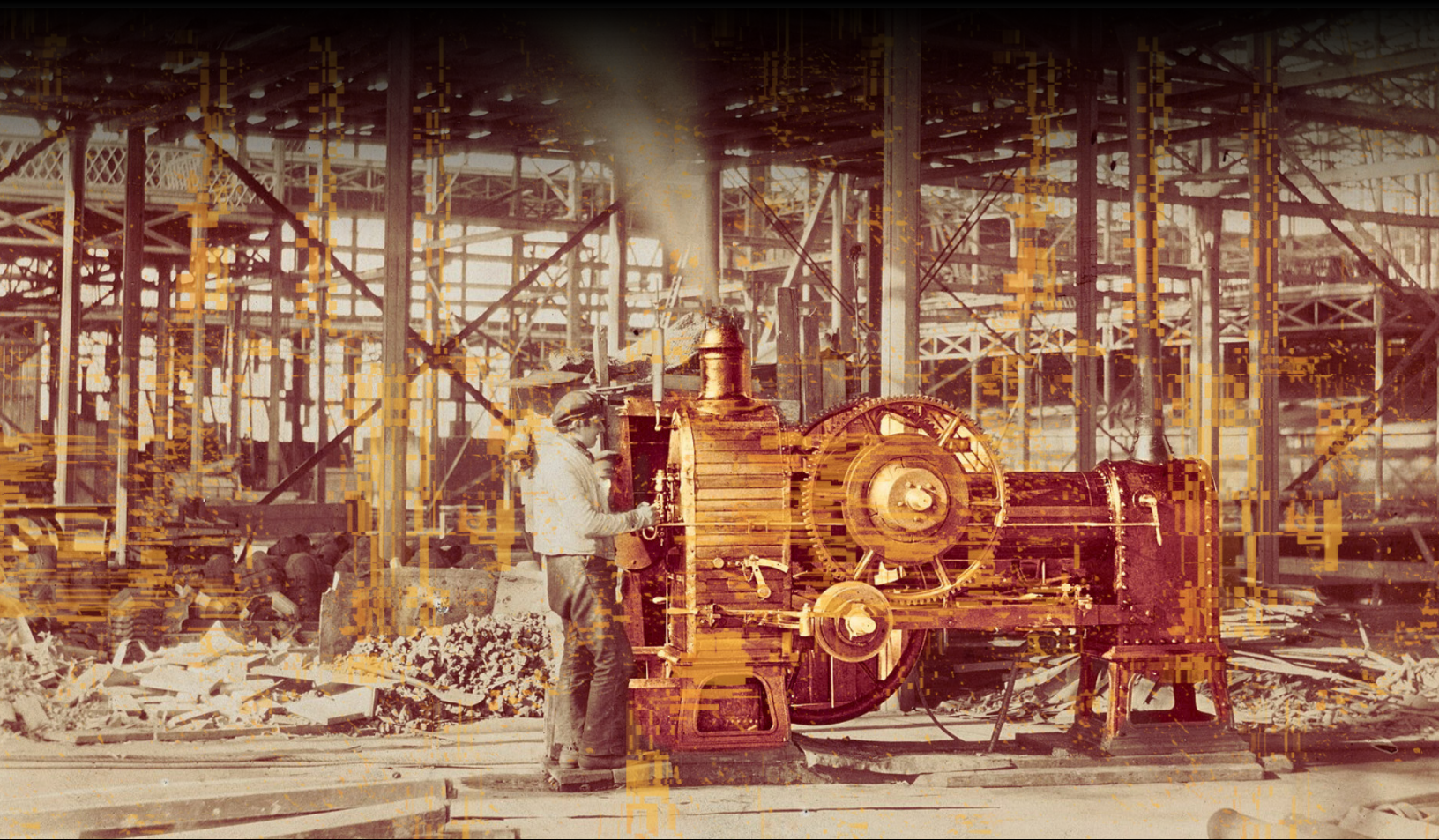


U C B E R K E L E Y
C E N T E R F O R L O N G - T E R M C Y B E R S E C U R I T Y



Transparency, Documentation, and Reporting Recommendations for General-Purpose AI Risk Management

**INCLUDING MAPPING OF GUIDANCE FROM THE NIST AI RISK MANAGEMENT
FRAMEWORK PLAYBOOK, THE G7 HIROSHIMA AI PROCESS REPORTING FRAMEWORK,
AND THE EU GPAI CODE OF PRACTICE**

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

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).

Transparency, Documentation, and Reporting Recommendations for General-Purpose AI Risk Management

INCLUDING MAPPING OF GUIDANCE FROM THE NIST AI RISK MANAGEMENT FRAMEWORK PLAYBOOK, THE G7 HIROSHIMA AI PROCESS REPORTING FRAMEWORK, AND THE EU GPAI CODE OF PRACTICE

**NADA MADKOUR[†] • JESSICA NEWMAN[†] • DEEPIKA RAMAN[†] • KRYSTAL JACKSON[†]
EVAN R. MURPHY[†] • CHARLOTTE YUAN[†]**

[†] AI Security Initiative, Center for Long-Term Cybersecurity, UC Berkeley

All affiliations listed are either current, or were during main contributions to this work or a previous version.

Version 1.0, April 2026

For the full General Purpose AI Risk-Management Standards Profile V1.2, and other supporting documents, see:

<https://cltc.berkeley.edu/publication/ai-risk-management-standards-profile-v1.2>



Contents

INTRODUCTION AND OBJECTIVES	4
Background	5
Industry Transparency Efforts	6
Academic Literature	7
Purpose and Scope	8
Intended Audience and Uses	8
Relationship with the GPAI Profile	8
Motivation and Importance	10
Limitations and Challenges	11
GUIDANCE	13
High-Priority Subcategories	13
Transparency, Documentation, and Reporting	
Mapping and Recommendations	15
Defining Roles and Responsibilities	15
Risk Documentation and Communication	16
External Feedback Processes	22
Risk Identification and Deployment Context	24
Risk Tolerance Determination	25
Likelihood and Impact Estimation	26
Risk Evaluation	27
Risk Tracking	30
Go/No-Go Decision	31
Risk Responses	32
Responses to Emergent Risk	35
Decommissioning Mechanisms	36
Post-Deployment Monitoring	37

TRANSPARENCY, DOCUMENTATION, AND REPORTING RECOMMENDATIONS
FOR GENERAL-PURPOSE AI RISK MANAGEMENT

GLOSSARY	40
Key Terms and Definitions	40
APPENDICES	41
Appendix 1: Unmapped Questions from the HAIP Reporting Framework	41
Appendix 2: Additional Resources	42
ACKNOWLEDGMENTS	43
REFERENCES	44

Introduction and Objectives

This document builds upon and complements the UC Berkeley General-Purpose AI (GPAI) Risk-Management Standards Profile, or “GPAI Profile” (Madkour et al. 2026), which provides risk-management practices and controls for identifying, analyzing, and mitigating risks of GPAI/foundation models. The purpose of this document is to provide AI model developers and deployers with recommendations related to transparency, documentation, and reporting that are in alignment with the GPAI Profile, and to provide a “crosswalk”¹ 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). The “Additional Recommendations” sections following the crosswalks includes guidance from priority resources such as California’s Transparency in Frontier Artificial Intelligence Act, SB53 (California Legislature 2025), and New York’s Responsible AI Safety and Education (RAISE) Act (New York State Senate, 2026).

By mapping existing transparency frameworks and providing additional guidance where gaps exist, this work offers AI developers, deployers, and policymakers a consolidated reference for transparency guidance that reflects current best practices, helping to cultivate stakeholder trust and support and responsible AI development and deployment. Robust transparency, documentation, and reporting of GPAI risk-management practices and policies complement transparency efforts that are common across many companies, including disclosure of financial performance and operations, as well as corporate social responsibility (CSR) reporting and environmental, social, and governance (ESG) reporting. Additionally, transparency has emerged as a critical governance tool to address regulatory gaps caused by over reliance on self-governance and voluntary commitments, and has been recognized as a powerful lever for effective AI risk management, particularly in cases where model developers may not be optimally positioned to implement the most effective risk mitigations.

¹ The “crosswalk” provides a structured mapping that links elements across these resources, highlighting how they correspond to one another, and helping enable alignment and interoperability.

BACKGROUND

Since 2023, there has been a sustained progression of government and multi-stakeholder efforts to address AI transparency. These efforts aim to provide visibility into AI risk management practices by sharing relevant information with stakeholders and affected parties. While initial efforts offered broad frameworks that incorporated transparency considerations, more recent developments have included dedicated frameworks and guidance that explicitly address AI transparency with greater depth and specificity. In 2023, early efforts included the **NIST AI RMF and Playbook** (NIST 2023a,b), and the **G7 HAIP Code of Conduct** (G7 2023), with each articulating guidance on documentation practices, information sharing, and transparency mechanisms. The momentum continued throughout 2024 with the release of the updated **OECD AI Principles**, the **NIST Generative AI Profile** (Autio et al. 2024), and the enactment of the **EU AI Act**. Each of these frameworks positioned transparency as a fundamental piece of effective AI governance, with the EU AI Act notably dedicating Article 50 to transparency obligations (EP 2024). The trend continued in 2025, governance bodies and multistakeholder initiatives began issuing dedicated, standalone guidance and regulatory requirements for AI transparency, most notably: the **EU GPAI Code of Practice** Transparency chapter (and the emphasis on transparency in the Safety and Security chapter) (EC 2025a), of which Anthropic, Amazon, Google, OpenAI, Microsoft, and many other AI companies are official signatories (EC 2025a); the **G7 HAIP Reporting Framework** (OECD.AI 2025), which has elicited over 20 voluntary report submissions from leading AI developers including (OECD.AI n.d.); California’s **Transparency in Frontier Artificial Intelligence Act, SB53** (California Legislature 2025); and New York’s **Responsible AI Safety and Education (RAISE) Act** (New York State Senate 2026). Many organizations and individuals across academia, civil society, and industry also published notable AI transparency guidance over these years (see discussion below). Collectively, these initiatives underscore the importance of transparency as a critical prerequisite to accountability and a foundational component of robust AI risk management.

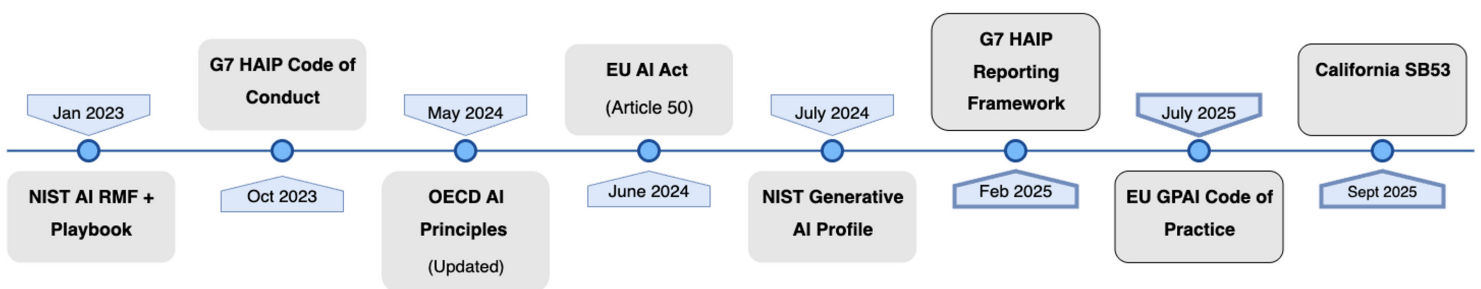


Figure 1. Government Transparency Initiatives Timeline

Industry Transparency Efforts

Industry transparency mechanisms include several commonly adopted practices: model/system cards documenting evaluation results and limitations (e.g., Anthropic 2025a, OpenAI 2025b, Meta 2025, Google 2025a); safety frameworks outlining risk management practices (e.g., Anthropic 2026, OpenAI 2025c, Google 2025b, Microsoft 2025a); centralized transparency platforms providing consolidated information and resources (e.g., OpenAI n.d., Anthropic n.d.b); organizational responsible AI transparency reports (Microsoft 2025b, Google 2025c); technical documentation for developers; explanatory guidance for users and stakeholders; and contractual B2B documentation (OECD 2025a).

A notable industry contribution is Anthropic’s proposed Frontier Model Transparency Framework (Anthropic 2025d), which is particularly notable due to the lack of observed similar efforts from other frontier model developers. The framework presents minimum transparency requirements, and includes requirements specific to pre-deployment. The framework also mandates the development of a Secure Development Framework (SDF) incorporating minimum standards, including assessment plans, whistleblower protections, processes for SDF modification, and retention of SDF documentation for five years. Additionally, this framework requires publication of a system card for each model deployment, or upon the addition of substantial capabilities to deployed models. The framework acknowledges the tension between transparency and security by permitting redaction of information that could materially compromise public safety. The framework echoes objectives in broader regulatory instruments such as the EU GPAI CoP (EC 2025a), California’s recently enacted law, SB53 (California Legislature 2025), and New York’s Responsible AI Safety and Education (RAISE) Act (New York State Senate 2026).

These efforts demonstrate the growing recognition of the importance of transparency for both external (e.g., users, affected communities) and internal (e.g., employees, management) stakeholders. However, closer examination reveals limitations that may constrain their effectiveness. An analysis by OECD (2025a) of responses to the Hiroshima AI Process Reporting Framework from 20 diverse organizations revealed significant variations in training data transparency practices, and affirms that disclosing training data information continues to pose challenges, as some organizations reserve detailed information for private contractual agreements. There was also variation in practices for updating documents, with some organizations providing versioned documents with accompanied change logs, and others only updating documentation for major changes in the model or new releases (OECD 2025a). These findings illuminate both the progress and challenges in industry transparency practices. While

TRANSPARENCY, DOCUMENTATION, AND REPORTING RECOMMENDATIONS FOR GENERAL-PURPOSE AI RISK MANAGEMENT

voluntary initiatives demonstrate intent to advance transparency, the variability in practices highlights the value of more standardized guidance.

Academic Literature

Researchers from academia and civil society have examined the state of AI transparency practices and proposed methodologies and frameworks to improve them, with many focusing on effectively communicating AI capabilities, limitations, and risks to a diverse set of stakeholders. Many of these efforts establish key principles for effective transparency, such as those proposed by Bommasani et al. (2024b) (i.e., centralization, structure, contextualization, independent specification, standardization, and methodologies),² and Chmielinski et al. (2024) (i.e., comparable, legible, actionable, and robust).

Other scholars have highlighted critical documentation artifact design considerations, including determining the appropriate degree of customization within transparency artifacts, the level of detail disclosed, and the extent of process automation (Winecoff and Bogen 2024). Despite the variance in methodology, these works consistently identify transparency as a necessary condition for facilitating accountability in AI systems (Bommasani et al. 2024a,b, Howell and Ifayemi 2024, Winecoff and Bogen 2024, Chmielinski et al. 2024, Lucaj et al. 2025). Additionally, researchers have emphasized the need for harmonized and standardized documentation practices to enhance interoperability and practical effectiveness (Crichton et al. 2025), and many have acknowledged that the lack of such standardization remains a persistent challenge in the field (Howell and Ifayemi 2024, Winecoff and Bogen 2024, Bommasani et al. 2024b, Chmielinski et al. 2024, Lucaj et al. 2025).

In empirical analysis, researchers evaluated the transparency of leading foundation developers using 100 indicators across 13 major dimensions of transparency. The results revealed that there have been significant improvements in transparency practices over time. Model developers scored on average 21 points (56%) higher in 2024 evaluations (58/100 indicators) compared to 2023 evaluations (37/100 indicators) (Bommasani et al. 2024a), signaling that measurable improvements in transparency are achievable. However, the 2025 update (Wan et al. 2025) showed a decline in transparency with the average score having fallen from 58 in 2024

² The six key principles highlighted by Bommasani et al. (2024b) are (1) centralization: consolidating information and transparency resources in a centralized location, (2) structure: reports follow a pre-defined structure that addresses specific queries and sets clear expectations, (3) contextualization: contextualizing information based on target stakeholders, (4) independent specification: independent specification of the underlying information to be included, (5) standardized: standardization of form and content reported, and (6) methodologies: clearly specified underlying methodologies for computing statistics.

TRANSPARENCY, DOCUMENTATION, AND REPORTING RECOMMENDATIONS FOR GENERAL-PURPOSE AI RISK MANAGEMENT

to 40 in 2025, highlighting the need for an increased focus on enhanced transparency from frontier model developers.

PURPOSE AND SCOPE

Intended Audience and Uses

The General-Purpose Artificial Intelligence Risk Management Standards Profile, or “GPAI Profile” (Madkour et al. 2026) provides risk-management practices or controls for identifying, analyzing, and mitigating risks of GPAI models. Transparency, documentation, and reporting are integrated throughout the GPAI Profile, and this document builds upon those initial recommendations with additional guidance, as well as an organized collection of best practices currently recommended across major regulations and initiatives. This work therefore is designed to complement and be used in conjunction with the GPAI Profile.

We intend for this document to be useful to developers of large-scale, state-of-the-art GPAI models, downstream developers of end-use applications or AI systems that build on a GPAI model, as well as third-party evaluators, policymakers, and regulators.

- **Model developers and deployers** can use this guidance as a consolidated reference and crosswalk to inform, design, and implement transparency, documentation, and reporting policies, practices, and procedures, ensuring compliance with existing frameworks and current best practices (e.g., G7 2023, EC 2025a, NIST 2023a).
- **Policymakers** can use this guidance as a consolidated reference to inform, design, and consider future policies, frameworks, or regulatory requirements around AI transparency and robust AI risk management.
- **Regulators** can use this guidance as an organized collection of best practices across major regulations and initiatives and as a tool for governance and enforcement.
- **Third-party evaluators** can use this guidance as a reference for a standardized collection of best practices and controls when assessing the transparency, documentation, and reporting of GPAI models.

Relationship with the GPAI Profile

The GPAI Profile primarily discusses transparency, documentation, and reporting in **Govern**
4.2: “Organizational teams document the risks and potential impacts of the AI technology they

TRANSPARENCY, DOCUMENTATION, AND REPORTING RECOMMENDATIONS FOR GENERAL-PURPOSE AI RISK MANAGEMENT

design, develop, deploy, evaluate, and use, and they communicate about the impacts more broadly.” Additional sub-categories incorporate transparency guidance and considerations to varying degrees, including:

- **Govern 1.4:** The risk management process and its outcomes are established through transparent policies, procedures, and other controls based on organizational risk priorities.
- **Govern 4.3:** Organizational practices are in place to enable AI testing, identification of incidents, and information sharing.
- **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.
- **Measure 2.8:** Risks associated with transparency and accountability – as identified in the Map function – are examined and documented.
- **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.
- **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.
- **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.

Additionally, throughout the GPAI Profile, we recommend documenting the processes used in considering risk mitigation controls, the evaluated options, and rationale for decisions made. We emphasize that while public-facing documentation should contain sufficient detail to enable meaningful transparency, information that could lead to exploitation by threat actors should be responsibly removed.

This document represents an opportunity to expand upon transparency guidance in the GPAI Profile by providing more comprehensive and detailed recommendations, along with a mapping of prominent transparency frameworks. Recognizing the critical role transparency plays in effective AI risk management and governance, we determined that a dedicated document would help address this foundational element with the appropriate level of depth. We have

TRANSPARENCY, DOCUMENTATION, AND REPORTING RECOMMENDATIONS FOR GENERAL-PURPOSE AI RISK MANAGEMENT

chosen to provide transparency recommendations aligned with our high-priority subcategories (rather than the full GPAI Profile) to deliver a focused set of recommendations on the stages of the risk management process where intervention is most critical.

Motivation and Importance

The motivation for this work is to provide a resource in alignment with the **NIST AI RMF characteristic of trustworthy AI, “accountable and transparent”** (NIST 2023a). This work aims to serve two essential functions: (1) consolidating transparency recommendations and requirements from prominent frameworks into a centralized resource, and (2) supplementing existing guidance where gaps may exist.

Transparency is an **indispensable component of effective AI risk management**, particularly in cases where model developers are not optimally positioned to implement the most effective risk mitigations. Managing AI risk often requires awareness and intervention from external stakeholders.³ Without sufficient visibility into model capabilities, limitations, and risks, external stakeholders cannot fulfill their roles in the AI risk management ecosystem. Additionally, even in contexts where developers are optimally positioned to mitigate risk, transparency remains essential to enable a multi-stakeholder defense-in-depth⁴ approach (Bengio et al. 2025).

In a regulatory landscape characterized by limited mandatory oversight, **AI governance remains predominantly dependent on self-governance** mechanisms and voluntary commitments, with varying outcomes (Wang et al. 2025). This reliance presents significant challenges to effective risk management, as voluntary frameworks intrinsically lack enforceable accountability structures, and provide no credible means for verifying organizational practices. Therefore, **transparency has emerged as a critical governance tool to address this regulatory gap**. Until sufficient mandatory regulations are established, transparency serves as the primary mechanism through which stakeholders can exercise oversight, as it establishes a structure for

³ For example, misuse of GPAI models for the generation of phishing email content may be better mitigated by platform operators and regulators. This does not exonerate model developers from managing these risks, but rather points to the need for a multi-stakeholder approach.

⁴ Defense in depth is a common technical approach that involves layering several controls and protective measures (Bengio et al. 2025). By establishing redundant safeguards, this approach ensures that the failure of any single protective measure does not result in complete system compromise or unmitigated risk.

TRANSPARENCY, DOCUMENTATION, AND REPORTING RECOMMENDATIONS FOR GENERAL-PURPOSE AI RISK MANAGEMENT

developer and deployer accountability. Additionally, transparency provides a means through which affected communities and other stakeholders can gain reassurance that AI systems are being developed, deployed, and managed with safety as a priority.

LIMITATIONS AND CHALLENGES

This document provides recommendations for best practices around transparency, such as what information to disclose and how to communicate it effectively. However, the recommendations **do not prescribe standards for the substantive content being reported**. For example, while we offer recommendations on what aspects of model evaluations to report, we do not provide guidance on how to perform rigorous evaluations. Consequently, adherence to these transparency recommendations supports alignment with disclosure and documentation best practices, but does not guarantee the underlying activities meet best practice standards. It is important to remember that **transparency in and of itself does not mean GPAI models will be safe, or that AI companies (i.e., model developers) will be held accountable for unsafe practices**.

Defining “**meaningful transparency**” varies across stakeholder groups, contexts, and use cases. There is significant variation in transparency definitions and requirements across different jurisdictions (Gyevnar et al. 2023, Matulionyte 2023). A uniform documentation approach cannot adequately address the needs of diverse stakeholder groups (Winecoff and Bogen 2024). For example, information that enables informed decision-making for users differs from information needed for downstream deployers to responsibly integrate a model. While we often indicate the type of transparency to which our recommendations correspond (e.g., reports to private stakeholders such as oversight bodies, or public documentation for users and affected communities), it is challenging to balance between making guidelines specific enough and ensuring they remain relevant across diverse stakeholders and contexts.

Another challenge is the **tension between transparency and security**. While information sharing can promote accountability and equity, excessive disclosure can introduce security and safety risks (Solaiman 2023). For example, the EU GPAI Code of Practice (CoP) emphasizes that publications should remove vulnerable information that may undermine security measures (EC 2025a). Our recommendations seek to limit prescriptive guidance and allow for necessary nuance when sensitive information is involved.

TRANSPARENCY, DOCUMENTATION, AND REPORTING RECOMMENDATIONS FOR GENERAL-PURPOSE AI RISK MANAGEMENT

While transparency in model evaluations is a critical accountability mechanism, fundamental **limitations in current understanding of advanced AI systems** must be acknowledged. Even with comprehensive access to models, evaluation results, and internal processes, developers and third-party evaluators often face knowledge gaps regarding how models function and what emergent capabilities may arise under specific conditions. These limitations suggest that transparency is insufficient as a standalone governance mechanism. Developers must explicitly acknowledge and report any gaps in model understanding within documentation, clearly distinguish between verified model behaviors and areas of uncertainty, and communicate the scope and limitations of evaluation coverage.

Guidance

HIGH-PRIORITY SUBCATEGORIES

Our General-Purpose AI Risk-Management Standards Profile, or “GPAI Profile” is part of a growing body of resources intended to identify and mitigate the risks of AI systems, which introduce novel privacy, security, equity, and malicious use concerns. Large-scale, cutting-edge GPAI models have the potential to behave unpredictably, manipulate or deceive humans in harmful ways, or lead to severe or catastrophic consequences. The GPAI Profile aims to ensure that developers of such systems take appropriate measures to anticipate and plan for a wide range of potential harms, from racial bias and environmental harms to destruction of critical infrastructure and degradation of democratic institutions.

Within the GPAI Profile, high-priority subcategories have been chosen based on baseline or minimum expectations for AI risk management. For more on high-priority risk-management steps and corresponding GPAI Profile guidance sections, see Madkour et al. (2026).

1. **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.
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.
3. **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.
4. **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.
5. **Map 1.5:** Organizational risk tolerances are determined and documented.
6. **Map 5.1:** Likelihood and magnitude of each identified impact (both potentially beneficial

TRANSPARENCY, DOCUMENTATION, AND REPORTING RECOMMENDATIONS
FOR GENERAL-PURPOSE AI RISK MANAGEMENT

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.

7. **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.
8. **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.
9. **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.
10. **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.
11. **Manage 2.3:** Procedures are followed to respond to and recover from a previously unknown risk when it is identified.
12. **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.
13. **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.

TRANSPARENCY, DOCUMENTATION, AND REPORTING MAPPING AND RECOMMENDATIONS

This section maps the transparency, documentation, and reporting recommendations found in the NIST AI RMF Playbook (NIST 2023b), the EU General-Purpose AI Code of Practice (GPAI CoP) (EC 2025a), and the G7 Hiroshima AI Process (HAIP) Transparency Report Questions (OECD.AI 2025) to each of our GPAI Profile high-priority subcategories (Madkour et al. 2026), followed by additional recommendations from leading resources such as California’s Transparency in Frontier Artificial Intelligence Act, SB53 (California Legislature 2025) and New York’s Responsible AI Safety and Education (RAISE) Act (New York State Senate 2026) where relevant.

Defining Roles and Responsibilities

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.

Govern 2.1 Mapped Recommendations

NIST AI RMF Playbook Transparency and Documentation	EU GPAI CoP (Safety and Security Chapter)	G7 (HAIP) Transparency Report Questions
<p>Organizations can document the following:</p> <ul style="list-style-type: none"> To what extent has the entity clarified the roles, responsibilities, and delegated authorities to relevant stakeholders? Who is ultimately responsible for the decisions of the AI and is this person aware of the intended uses and limitations of the analytic? Are the responsibilities of the personnel involved in the various AI governance processes clearly defined? What are the roles, responsibilities, and delegation of authorities of personnel involved in the design, development, deployment, assessment, and monitoring of the AI system? Did your organization implement accountability-based practices in data management and protection (e.g., the PDPA and OECD Privacy Principles)? 	<p>Measure 8.1: Signatories will clearly define responsibilities for managing the systemic risks [...] across all levels of the organisation [...]:</p> <ul style="list-style-type: none"> Systemic risk oversight: Overseeing the Signatories’ systemic risk assessment and mitigation processes and measures. Systemic risk ownership: Managing systemic risks stemming from Signatories’ models, including the systemic risk assessment and mitigation processes and measures, and managing the response to serious incidents. Systemic risk support and monitoring: Supporting and monitoring the Signatories’ systemic risk assessment and mitigation processes and measures. Systemic risk assurance: Providing internal and, as appropriate, external assurance about the adequacy of the Signatories’ systemic risk assessment and mitigation processes and measures to the management [...] or another suitable independent body (such as a council or board). <p>Signatories will allocate these responsibilities [...] across the following levels of their organisation:</p> <ul style="list-style-type: none"> The management body in its supervisory function or another suitable independent body (such as a council or board); The management body in its executive function; Relevant operational teams; If available, internal assurance providers (e.g., an internal audit function); and If available, external assurance providers (e.g., third-party auditors). 	<p>Section 4:</p> <ul style="list-style-type: none"> How has AI risk management been embedded in your organization governance framework? When and under what circumstances are policies updated? Are relevant staff trained on your organization’s governance policies and risk management practices? If so, how?

TRANSPARENCY, DOCUMENTATION, AND REPORTING RECOMMENDATIONS FOR GENERAL-PURPOSE AI RISK MANAGEMENT

Govern 2.1 Additional Recommendations

- Define and document roles and responsibilities for both internal and external AI actors:
 - AI actors may include data collectors, data processors, developers, modellers, system integrators, system operators, end-users, and stakeholders (OECD, 2022).
 - » Additional actors may include third-party entities, affected individuals and communities, members of the research community, and advocacy groups (NIST 2023a).
 - » Each of these actors may be involved in one or more of the AI system lifecycle stages (design, development, deployment, and operations and monitoring). For example, domain experts typically play a role in all four stages, while system integrators are involved in the deployment phase (NIST 2023a).
- Define and document roles and organizational policies for staff training on organizational governance policies, relevant laws, and organizational risk management practices:
 - These policies should include considerations around the different training and education needs of different types of staff, and mechanisms for acknowledgment.
 - CSA (2024) outlines potential roles and their descriptions within four domains: (1) management and strategy, (2) governance, risk, and compliance, (3) technical and security, and (4) operations and development. While illustrations of common roles can provide a helpful starting point, organizations are encouraged to define roles and responsibilities in a way that reflects their unique operational requirements.

For more on roles and responsibilities, see Govern 2.1 and Map 5.2 in Madkour et al. (2026).

Risk Documentation and Communication

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.*

Govern 4.2 Mapped Recommendations

NIST AI RMF Playbook Transparency and Documentation	EU GAPI CoP (Safety and Security, and Transparency Chapters)	G7 (HAIP) Transparency Report Questions
<p><i>Organizations can document the following:</i></p> <ul style="list-style-type: none"> • <i>How has the entity identified and mitigated potential impacts of bias in the data, including inequitable or discriminatory outcomes?</i> • <i>How has the entity documented the AI system’s data provenance, including sources, origins, transformations, augmentations, labels, dependencies, constraints, and metadata?</i> 	<p>Transparency Chapter: Measure 1.1 <i>Signatories, when placing a general-purpose AI model on the market, will have documented at least all the information referred to in the Model Documentation Form [...].⁵</i></p> <p>Safety and Security Chapter: Measure 1.1 <i>Signatories will create a state-of-the-art Framework [which] will contain a high-level description of implemented and planned processes and measures for systemic risk assessment and mitigation [...]. The Framework will contain:</i></p>	<p>Section 3:</p> <ul style="list-style-type: none"> • <i>Does your organization publish clear and understandable reports and/or technical documentation related to the capabilities, limitations, and domains of appropriate and inappropriate use of advanced AI systems?</i> <ul style="list-style-type: none"> ◦ <i>How often are such reports usually updated?</i> ◦ <i>How are new significant releases reflected in such reports?</i>

⁵ The model documentation form requires signatories to disclose: general information (e.g., model name, authenticity, release date, dependencies); model properties (e.g., model architecture, design specifications, input modalities, output modalities, model size); methods of distribution licenses (e.g., distribution channels, license); use (acceptable use policy, intended uses, system integrations, required hardware);

TRANSPARENCY, DOCUMENTATION, AND REPORTING RECOMMENDATIONS
FOR GENERAL-PURPOSE AI RISK MANAGEMENT

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.

Govern 4.2 Mapped Recommendations

NIST AI RMF Playbook Transparency and Documentation	EU GAPI CoP (Safety and Security, and Transparency Chapters)	G7 (HAIP) Transparency Report Questions
<ul style="list-style-type: none"> To what extent has the entity clearly defined technical specifications and requirements for the AI system? To what extent has the entity documented and communicated the AI system’s development, testing methodology, metrics, and performance outcomes? Have you documented and explained that machine errors may differ from human errors? 	<ul style="list-style-type: none"> A description and justification of the trigger points and their usage, at which the Signatories will conduct additional lighter-touch model evaluations along the entire model lifecycle; For the Signatories’ determination of whether systemic risk is considered acceptable: <ul style="list-style-type: none"> A description and justification of the systemic risk acceptance criteria, including the systemic risk tiers [...]; A high-level description of what safety and security mitigations Signatories would need to implement once each systemic risk tier is reached; For each systemic risk, estimates of timelines when Signatories reasonably foresee that they will have a model that exceeds the highest systemic risk tier [...]. Such estimates: <ul style="list-style-type: none"> May consist of time ranges or probability distributions; and May take into account aggregate forecasts, surveys, and other estimates produced with other providers. Further, such estimates will be supported by justifications, including underlying assumptions and uncertainties; and A description of [...] by what process input from external actors influences proceeding with the development, making available on the market, and/or use of the models. A description of how systemic risk responsibility is allocated for the processes by which systemic risk is assessed and mitigated; and A description of the process by which Signatories will update the Framework, including how they will determine that an updated Framework is confirmed. <p>Measure 1.3 Signatories will conduct an appropriate Framework assessment, if they have reasonable grounds to believe that the adequacy of their Framework and/or their adherence thereto has been or will be materially undermined, or every 12 months starting from their placing of the model on the market [...]. Examples of such grounds are:</p>	<ul style="list-style-type: none"> Which of the following information is included in your organization’s publicly available documentation: details and results of the evaluations conducted for potential safety, security, and societal risks including risks to the enjoyment of human rights; assessments of the model’s or system’s effects and risks to safety and society (such as those related to harmful bias, discrimination, threats to protection of privacy or personal data, fairness); results of red teaming or other testing conducted to evaluate the model’s/system’s fitness for moving beyond the development stage; capacities of a model/system and significant limitations in performance with implications for appropriate use domains; other technical documentation and instructions for use if relevant. Does your organization demonstrate transparency related to advanced AI systems through any other methods? Does your organization disclose privacy policies addressing the use of personal data, user prompts, and/or the outputs of advanced AI systems? Does your organization provide information about the sources of data used for the training of advanced AI systems, as appropriate, including information related to the sourcing of data annotation and enrichment?

training process (e.g., design specifications, decision rationale); TEVV data (e.g., modality/type, data provenance, data gathering and selection, number of data points, scope and main characteristics, curation methodologies, measures to detect unusability and bias); computation resources (e.g., training time, training compute, measurement methodology); energy consumption (e.g., energy used for training, measurement methodology). See the EU GPAI Code of Practice (EC 2025a) Transparency chapter for the full model documentation form.

TRANSPARENCY, DOCUMENTATION, AND REPORTING RECOMMENDATIONS
FOR GENERAL-PURPOSE AI RISK MANAGEMENT

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.

Govern 4.2 Mapped Recommendations

NIST AI RMF Playbook Transparency and Documentation	EU GAPI CoP (Safety and Security, and Transparency Chapters)	G7 (HAIP) Transparency Report Questions
	<ul style="list-style-type: none"> • <i>How the Signatories develop models will change materially, which can be reasonably foreseen to lead to the systemic risks stemming from at least one of their models not being acceptable;</i> • <i>Serious incidents and/or near misses involving their models or similar models that are likely to indicate that the systemic risks stemming from at least one of their models are not acceptable have occurred; and/or</i> • <i>The systemic risks stemming from at least one of their models have changed or are likely to change materially, e.g., safety and/or security mitigations have become or are likely to become materially less effective, or at least one of their models has developed or is likely to develop materially changed capabilities and/or propensities.</i> <p>Measure 9.2 <i>Keep track of, document, and report at least the following information to the best of their knowledge, redacted to the extent necessary:</i></p> <ul style="list-style-type: none"> • <i>The start and end dates of the serious incident, or best approximations;</i> • <i>The resulting harm and the victim or affected group;</i> • <i>The chain of events that (directly or indirectly) led to the serious incident;</i> • <i>The model involved in the serious incident;</i> • <i>A description of material available setting out the model's involvement in the serious incident;</i> • <i>Response to the serious incident;</i> • <i>A root cause analysis with a description of the model's outputs that (directly or indirectly) led to the serious incident and the factors that contributed to their generation, including the inputs used and any failures or circumventions of systemic risk mitigations; and</i> • <i>Any patterns detected during post-market monitoring that can reasonably be assumed to be connected to the serious incident.</i> 	<p>Section 4:</p> <ul style="list-style-type: none"> • <i>How does your organization share relevant information about vulnerabilities, incidents, emerging risks, and misuse with others?</i> • <i>Does your organization share information, as appropriate, with relevant other stakeholders regarding advanced AI system incidents? If so, how?</i> • <i>Does your organization share and report incident-related information publicly?</i> • <i>Does your organization communicate its risk management policies and practices with users and/or the public? If so, how?</i> • <i>How does your organization share research and best practices on addressing or managing risk?</i>

TRANSPARENCY, DOCUMENTATION, AND REPORTING RECOMMENDATIONS
FOR GENERAL-PURPOSE AI RISK MANAGEMENT

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.*

Govern 4.2 Mapped Recommendations

NIST AI RMF Playbook Transparency and Documentation	EU GAPI CoP (Safety and Security, and Transparency Chapters)	G7 (HAIP) Transparency Report Questions
	<p>Measure 9.4: <i>Signatories will keep documentation of all relevant information gathered in adhering to this Commitment for at least five years from the date of the documentation or the date of the serious incident, whichever is later [...].</i></p> <p>Measure 10.1: <i>Signatories will draw up and keep up-to-date the following information for the purpose of providing it to the AI Office upon request:</i></p> <ul style="list-style-type: none"> • <i>A detailed description of the model's architecture;</i> • <i>A detailed description of how the model is integrated into AI systems, explaining how software components build or feed into each other and integrate into the overall processing, insofar as the Signatory is aware of such information;</i> • <i>A detailed description of the model evaluations conducted pursuant to this Chapter, including their results and strategies; and</i> • <i>A detailed description of the safety mitigations implemented (pursuant to Commitment 5).</i> <p><i>Documentation will be retained at least 10 years after the model has been placed on the market. Further, Signatories will keep track of the following information, to the extent it is not already covered by the first paragraph, for the purpose of evidencing adherence to this Chapter to the AI Office upon request:</i></p> <ul style="list-style-type: none"> • <i>Their processes, measures, and key decisions that form part of their systemic risk assessment and mitigation; and</i> • <i>Justifications for choices of a particular best practice, state-of-the-art, or other more innovative process or measure if a Signatory relies upon it for adherence to this Chapter.</i> <p>Measure 10.2 <i>If and insofar as necessary to assess and/or mitigate systemic risks, Signatories will publish [...] a summarised version of their Framework and Model Report(s), and updates thereof, with removals to not undermine the effectiveness of safety and/or security mitigations and to protect sensitive commercial information.</i></p> <ul style="list-style-type: none"> • <i>For Model Reports, such publication will include high-level descriptions of the systemic risk assessment results and the safety and security mitigations implemented.</i> 	

TRANSPARENCY, DOCUMENTATION, AND REPORTING RECOMMENDATIONS FOR GENERAL-PURPOSE AI RISK MANAGEMENT

Govern 4.2 Additional Recommendations

- Establish standardized, tiered disclosure protocols and policies that provide differentiated levels of detail to various stakeholders.
 - Information sharing across stakeholders should be done while maintaining consistent timing, methodology transparency, and clear articulation of limitations and uncertainties across all reporting levels.
 - Establish secure, structured information-sharing mechanisms with other developers to collectively identify emerging safety and security risks, and share anonymized failure modes and near-miss incidents. Information sharing may be done with appropriate safeguards to protect competitive information and prevent inadvertent disclosure of sensitive capabilities or vulnerabilities. Information sharing should include:
 - » Vulnerabilities;
 - » Emerging risks;
 - » Incidents and near misses;
 - » Misuse patterns; and
 - » Research and best practices on risk mitigations and risk management.
 - Establish secure, structured information-sharing mechanisms with oversight and regulatory bodies to report incidents and near misses:
 - » Disclose critical safety incidents within a pre-defined period of time (e.g., 15 days per the California Legislature 2025, and 72 hours per New York State Senate 2026).
 - » Disclose critical safety incidents that pose an imminent threat to physical health or death within 24 hours (California Legislature 2025).
- Establish and make available, and easily accessible, relevant public-facing policies, including:
 - Privacy policies;
 - Acceptable use policies (AUPs);
 - Prohibited uses; and
 - Data use and retention policies.
- Consider making all management policies, frameworks, and practices publicly available through accessible, centralized platforms. For example, Anthropic’s Transparency Hub provides stakeholders with a dedicated portal that consolidates and presents the company’s policies in a user-friendly format (Anthropic n.d.b).
- Report incidents and near misses to public databases or repositories,⁶ such as the AI Incident Database (AIID n.d.), and others.
 - When reporting incidents, include the following information:
 - » Date of the incident (California Legislature 2025, New York State Senate 2026);
 - » A description of the incident (California Legislature 2025, New York State Senate 2026);
 - » The reasons the incident qualifies as a safety incident (California Legislature 2025, New York State Senate 2026); and
 - » Whether or not this was an incident associated with internal use of the model (California Legislature 2025, New York State Senate 2026).
- AI developers should retain an unredacted copy and publish a version of their safety frameworks with the necessary redactions (California Legislature 2025, New York State Senate 2026).
- AI developers should publish public versions of their model evaluation reports (i.e., model cards and system cards⁷), with information on and clear explanations of:
 - Release date of the model (California Legislature 2025);
 - Training data sources and annotation;

6 Other examples of risk registers and incident databases include: AI Incident Database (AIID n.d.), ATLAS AI Incidents (MITRE n.d.a), MIT AI Incident Tracker (MIT 2025a), MIT AI Risk Repository (MIT 2025b), AI Risk Database (MITRE n.d.b), AI Vulnerability Database (AVID n.d.).

7 For examples of model/system cards see, GPT-5 (OpenAI 2025a) Claude Opus 4.1 (Anthropic 2025a), and Llama 4 (Meta n.d.).

TRANSPARENCY, DOCUMENTATION, AND REPORTING RECOMMENDATIONS FOR GENERAL-PURPOSE AI RISK MANAGEMENT

Govern 4.2 Additional Recommendations

- Model safety, security, and risk evaluation methodologies and results;
- Methodology and results of systemic risk evaluations (e.g., evaluations of harmful bias, discrimination, threats to protection of privacy or personal data, and fairness);
- A description of how results were translated into risk scores or risk levels;
- Methodology and results of other types of testing conducted (e.g., red teaming);
- Implemented security mitigations and controls;
- Justification for moving beyond the development phase;
- Capacities and significant limitations of the model;
- Implications of use; and
- Other instructions and technical documentation, if relevant.
- Each model release should be accompanied by a new report or a report addendum.
- Consider making available insights into usage analysis. For example, Anthropic has published its Economic Index (Anthropic n.d.), a tool that helps understand Claude’s usage across U.S. states, jobs, and task categories.
- Consider publishing research results and best practices for risk management and mitigation.
 - For example, Anthropic and Google DeepMind provide public research results (Anthropic n.d.a, Google DeepMind n.d.).

(For more on model evaluation transparency, documentation, and reporting, see Measure 1.1 in this document. For more on risk tiers and risk tolerance, transparency, documentation, and reporting, see Map 1.5 in this document. For more on risk impact and likelihood estimating, see Map 5.1 in this document.)

TRANSPARENCY, DOCUMENTATION, AND REPORTING RECOMMENDATIONS
FOR GENERAL-PURPOSE AI RISK MANAGEMENT

External Feedback Processes

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.

Govern 5.1 Mapped Recommendations

NIST AI RMF Playbook Transparency and Documentation	EU GPAI CoP (Safety and Security Chapter)	G7 (HAIP) Transparency Report Questions
<p>Organizations can document the following:</p> <ul style="list-style-type: none"> • 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 has the entity clarified the roles, responsibilities, and delegated authorities to relevant stakeholders? • How easily accessible and current is the information available to external stakeholders? • 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. 	<p>Measure 3.1: Gather model-independent information relevant to the systemic risk [...] with varying degrees of breadth and depth appropriate for the systemic risk, using methods such as:</p> <ul style="list-style-type: none"> • Expert interviews and/or panels; and • Lay interviews, surveys, community consultations, or other participatory research methods. <p>Measure 3.5: The following are examples of methods for the collection of information:</p> <ul style="list-style-type: none"> • Collecting end-user feedback; • Providing (anonymous) reporting channels; • Providing (serious) incident reporting forms; and • Providing bug bounties. <p>Measure 7.4: Signatories will provide in the Model Report any available reports (e.g., via valid hyperlinks) from:</p> <ul style="list-style-type: none"> • Independent external evaluators involved in model evaluations; and • Security reviews undertaken by an independent external party. <p>Measure 8.3: Signatories will promote a healthy risk culture by providing:</p> <ul style="list-style-type: none"> • Anonymous surveys; and • Internal reporting channels. 	<p>Section 1:</p> <ul style="list-style-type: none"> • Does your organization make vulnerability and incident reporting mechanisms accessible to a diverse set of stakeholders? • Does your organization have incentive programs for the responsible disclosure of risks, incidents, and vulnerabilities? • How does your organization collaborate with relevant stakeholders across sectors to assess and adopt risk mitigation measures to address risks, in particular systemic risks? <p>Section 4:</p> <ul style="list-style-type: none"> • Does your organization use international technical standards or best practices for AI risk management and governance policies? • Are steps taken to address reported incidents documented and maintained internally? If so, how? <p>Section 7:</p> <ul style="list-style-type: none"> • Does your organization collaborate with civil society and community groups to identify and develop AI solutions in support of the UN Sustainable Development Goals and to address the world's greatest challenges? Please provide examples.

TRANSPARENCY, DOCUMENTATION, AND REPORTING RECOMMENDATIONS
FOR GENERAL-PURPOSE AI RISK MANAGEMENT

Govern 5.1 Additional Recommendations

- Establish policies and procedures for easily accessible third-party reporting mechanisms that include:
 - Reporting forms for serious incidents (e.g., see the EU reporting template for serious AI incidents, EC 2025a);
 - Anonymous reporting channels;
 - Protections for good-faith disclosures (Longpre et al. 2024);
 - Frequent communication with affected stakeholders;
 - Bi-directional feedback mechanisms that facilitate active engagement and an iterative exchange of information;
 - A mechanism to allow user/external communication with the model developer (California Legislature 2025); and
 - Incentive programs for responsible disclosure of risks, vulnerabilities, and misaligned model behavior:
 - » Vulnerability disclosure programs (e.g., OpenAI 2025b);
 - » Bug bounty programs (e.g., Anthropic 2025c);
 - » Misalignment Bounty programs (e.g., Palisade Research n.d.);
 - » Legal protections for good-faith reporting; and
 - » Coordinated flaw disclosure programs (e.g., Longpre et al. 2025).
- To effectively address and document reported incidents, establish incident response policies that include:
 - Practices for documentation of relevant incident information, such as:
 - » Start and end dates (or an approximation);
 - » Resulting harms;
 - » Affected persons or groups;
 - » Incident chain of events and root cause analysis;
 - » Information on the mode involved (e.g., version); and
 - » Responses to the incident.
 - Monitoring systems for reporting, tracking, and analyzing incidents;
 - Practices for external reporting and notification of incidents to public databases — such as the AI Incident Database (AIID n.d.), and the MIT AI Incident Tracker (MIT 2025a) — or oversight bodies; and
 - Roles, responsibilities, and points of contact for incident response and stakeholder communication.
- Establish whistleblower protection policies and employee protections such as those set forth in California Legislature (2025).⁸
 - Contracts should never prevent employees from making disclosures protected under the whistleblower protection policies.
 - Permit covered employees to use the whistleblower hotline for reporting.
 - Ensure that employees are notified and aware of their rights and responsibilities pursuant to any and all whistleblower protection policies:
 - » Re-notify employees at least once a year.
 - » Make the policies accessible and available to covered employees at all times.
 - Provide an internal anonymous reporting option for employees.
 - Provide at least monthly updates to the disclosing party on the status of the reported issue.
 - Share disclosures and responses to organizational directors at least once per quarter.
- In addition to seeking out feedback from external stakeholders, consider, prioritize, and integrate technical standards and best practices (international and local. E.g., ISO/IEC 42001:2023 (ISO 2023)).
- Develop practices and policies for partnering with third-party evaluators and auditors.
- Develop practices and policies for collaboration with civil society and community groups.
- Develop practices and policies for responsible information sharing and partnering with other model developers.

⁸ For more on whistleblower protections, see Sec. 4, Chapter 5.1 of the Transparency in Frontier Artificial Intelligence Act, S.B. 53 (California Legislature 2025).

TRANSPARENCY, DOCUMENTATION, AND REPORTING RECOMMENDATIONS
FOR GENERAL-PURPOSE AI RISK MANAGEMENT

Risk Identification and Deployment Context

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.*

Map 1.1 Mapped Recommendations

NIST AI RMF Playbook Transparency and Documentation	EU GPAI CoP (Safety and Security Chapter)	G7 (HAIP) Transparency Report Questions
<p>Organizations can document the following:</p> <ul style="list-style-type: none"> • To what extent is the output of each component appropriate for the operational context? • Which AI actors are responsible for the decisions of the AI and is this person aware of the intended uses and limitations of the analytic? • Which AI actors are responsible for maintaining, re-verifying, monitoring, and updating this AI once deployed? • Who is the person(s) accountable for the ethical considerations across the AI lifecycle? 	<p>Measure 2.1: <i>Signatories will identify the systemic risks obtained through the following process:</i></p> <ul style="list-style-type: none"> • Compiling a list of risks that could stem from the model, taking into account: <ul style="list-style-type: none"> ◦ Model-independent information; ◦ Relevant information about the model and similar models, including information from post-market monitoring, and information about serious incidents and near misses; and ◦ Any other relevant information communicated directly or via public releases to the Signatory. • Analysing relevant characteristics of the risks compiled pursuant to point, such as their nature and sources. • Identifying the systemic risks stemming from the model. <p>Measure 2.2: <i>Signatories will develop appropriate systemic risk scenarios, including regarding the number and level of detail of these systemic risk scenarios, for each identified systemic risk (pursuant to Measure 2.1).</i></p>	<p>Section 1:</p> <ul style="list-style-type: none"> • What practices does your organization use to identify and evaluate risks such as vulnerabilities, incidents, emerging risks, and misuse, throughout the AI lifecycle? • Does your organization contribute to the development of and/or use international technical standards or best practices for the identification, assessment, and evaluation of risks? <p>Section 3:</p> <ul style="list-style-type: none"> • Does your organization disclose privacy policies addressing the use of personal data, user prompts, and/or the outputs of advanced AI systems?

Map 1.1 Discussion and Additional Considerations

- Establish and make publicly available, and easily accessible, relevant public-facing documentation that cover:
 - Privacy policies;
 - Acceptable use policies (AUPs), that cover elements such as those included in SB53 (California Legislature 2025):
 - » Languages supported by the model;
 - » Modalities of output;
 - » Intended uses; and
 - » Restrictions or conditions of use.
 - Prohibited uses;
 - Data use and retention policies; and
 - Technical guidance and instructions as needed.

TRANSPARENCY, DOCUMENTATION, AND REPORTING RECOMMENDATIONS
FOR GENERAL-PURPOSE AI RISK MANAGEMENT

Risk Tolerance Determination

Map 1.5: <i>Organizational risk tolerances are determined and documented.</i>		
Map 1.5 Mapped Recommendations		
NIST AI RMF Playbook Transparency and Documentation	EU GPAI CoP (Safety and Security Chapter)	G7 (HAIP) Transparency Report Questions
<p>Organizations can document the following:</p> <ul style="list-style-type: none"> • Which existing regulations and guidelines apply, and the entity has followed, in the development of system risk tolerances? • 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? 	<p>Measure 4.1: <i>Signatories will describe and justify [...] how they will determine whether the systemic risks [...] are acceptable. To do so, Signatories will:</i></p> <ul style="list-style-type: none"> • for each identified systemic risk, at least: <ul style="list-style-type: none"> ◦ Define appropriate systemic risk tiers that: <ul style="list-style-type: none"> » Are defined in terms of model capabilities, and may additionally incorporate model propensities, risk estimates, and/or other suitable metrics; » Are measurable; and » Comprise at least one systemic risk tier that has not been reached by the model ◦ Define other appropriate systemic risk acceptance criteria, if systemic risk tiers are not suitable for the systemic risk and the systemic risk is not a specified systemic risk. • Describe how [...] these tiers and/or other criteria to determine whether each identified systemic risk and the overall systemic risk are acceptable; and • Justify how the use [...] ensures that each identified systemic risk and the overall systemic risk are acceptable. <p>Signatories will apply the systemic risk acceptance criteria to each identified systemic risk, incorporating a safety margin [...] to determine whether each identified systemic risk and the overall systemic risk are acceptable. This acceptance determination will take into account at least the information gathered via systemic risk identification and analysis.</p> <p>The safety margin will:</p> <ul style="list-style-type: none"> • Be appropriate for the systemic risk; and • Take into account potential limitations, changes, and uncertainties of: <ul style="list-style-type: none"> ◦ Systemic risk sources (e.g., capability improvements after the time of assessment); ◦ Systemic risk assessments (e.g., under-elicitation of model evaluations or historical accuracy of similar assessments); and ◦ The effectiveness of safety and security mitigations (e.g., mitigations being circumvented, deactivated, or subverted). 	<p>Section 1:</p> <ul style="list-style-type: none"> • How does your organization define and/or classify different types of risks related to AI, such as unreasonable risks?

TRANSPARENCY, DOCUMENTATION, AND REPORTING RECOMMENDATIONS
FOR GENERAL-PURPOSE AI RISK MANAGEMENT

Map 1.5 Additional Recommendations

- Safety frameworks should clearly articulate and justify risk tiers and risk acceptance criteria. This should include:
 - Quantitative metrics for measuring risk tiers where possible;
 - Several risk tiers below the intolerable tier, to allow for appropriate response time;
 - A description of the methodologies and logic used to measure and define the risk tiers; and
 - A description of how thresholds for determining whether the model possesses intolerable capabilities were defined and assessed (California Legislature 2025).
- Engage the participation of independent actors, including governments, academic institutions, public-interest research organizations, standard-setting bodies, and civil society groups in the risk-tier identification, verification, and operationalization process (Newman et al. 2025).

Likelihood and Impact Estimation

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.

Map 5.1 Mapped Recommendations

NIST AI RMF Playbook Transparency and Documentation	EU GPAI CoP (Safety and Security Chapter)	G7 (HAIP) Transparency Report Questions
<p>Organizations can document the following:</p> <ul style="list-style-type: none"> • Which population(s) does the AI system impact? • What assessments has the entity conducted on trustworthiness characteristics for example data security and privacy impacts associated with the AI system? • Can the AI system be tested by independent third parties? 	<p>Measure 3.4: Signatories will estimate the probability and severity of harm for the systemic risk.</p> <p>Signatories will use at least state-of-the-art risk estimation methods and take into account at least the information gathered [during systemic risk identification], [from] this Commitment, and [from serious incident reporting]. Estimates of systemic risk will be expressed as a risk score, risk matrix, probability distribution, or in other adequate formats, and may be quantitative, semi-quantitative, and/or qualitative. Examples of such estimates of systemic risks are:</p> <ul style="list-style-type: none"> • A qualitative systemic risk score (e.g., “moderate” or “critical”); • A qualitative systemic risk matrix (e.g., “probability: unlikely” x “impact: high”); and/or • A quantitative systemic risk matrix (e.g. “X-Y%” x “X-Y EUR damage”). 	<p>Section 1:</p> <ul style="list-style-type: none"> • Does your organization use incident reports, including reports shared by other organizations, to help identify risks? • What practices does your organization use to identify and evaluate risks such as vulnerabilities, incidents, emerging risks and misuse, throughout the AI lifecycle? Are quantitative and/or qualitative risk evaluation metrics used and if yes, with what caveats?

Map 5.1 Additional Recommendations

- Establish policies and procedures for risk estimation:
 - Take into consideration the information gathered in the risk identification process (see, Map 1.1).
 - Identify the appropriate combination of methods for estimating likelihood and magnitude for each identified risk type. Methods include:
 - » Risk matrices;
 - » Probability distributions;

TRANSPARENCY, DOCUMENTATION, AND REPORTING RECOMMENDATIONS
FOR GENERAL-PURPOSE AI RISK MANAGEMENT

Map 5.1 Additional Recommendations

- » Probabilistic risk assessments (Wisakanto et al. 2025);⁹ and
- » Risk modeling (Jackson et al. 2026)..
- Refer to incident data bases and risk repositories¹⁰ to help evaluate the likelihood of risk occurrence (e.g., AIID n.d., MIT 2025b).
- Include the participation of independent actors, including governments, academic institutions, public-interest research organizations, standard-setting bodies, and civil society groups in the risk estimation process (Newman et al. 2025).
- Document in detail the methodologies used to estimate the likelihood and impact of risks, and the logic behind risk score assignments.
- Documents any limitations or caveats of the chosen methods.
- Document in detail the risk estimation results for each model, and retain the documentation for an appropriate period of time.
- In safety frameworks, for each identified risk type, include descriptions and justifications for:
 - Risk likelihood and magnitude measurement methods; and
 - Risk likelihood and magnitude scoring logic.
- In model evaluation reports (i.e., model cards and system cards), for each identified risk type, include descriptions of:
 - Risk likelihood and magnitude measurement methods; and
 - Risk likelihood and magnitude estimation scores.

Risk Evaluation

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.

Measure 1.1 Mapped Recommendations

NIST AI RMF Playbook Transparency and Documentation	EU GPAI CoP (Safety and Security Chapter)	G7 (HAIP) Transparency Report Questions
<p>Organizations can document the following:</p> <ul style="list-style-type: none"> • How will the appropriate performance metrics, such as accuracy of the AI, be monitored after the AI is deployed? • What corrective actions has the entity taken to enhance the quality, accuracy, reliability, and representativeness of the data? • Are there recommended data splits or evaluation measures? (e.g., training, development, testing; accuracy/AUC) 	<p>Measure 3.2: Signatories will conduct at least state-of-the-art model evaluations in the modalities relevant to the systemic risk to assess the model’s capabilities, propensities, affordances, and/or effects.</p> <p>Signatories will ensure that such model evaluations are designed and conducted using methods that are appropriate for the model and the systemic risk, and include open-ended testing of the model to improve the understanding of the systemic risk, with a view to identifying unexpected behaviours, capability boundaries, or emergent properties. Examples of model evaluation methods are: Q&A sets, task-based evaluations, benchmarks, red-teaming and other methods of adversarial testing, human uplift studies, model organisms, simulations, and/or proxy evaluations</p>	<p>Section 1:</p> <ul style="list-style-type: none"> • Describe how your organization conducts testing (e.g., red-teaming) to evaluate the model’s/system’s fitness for moving beyond the development stage? • Are quantitative and/or qualitative risk evaluation metrics used, and if yes, with what caveats? • Is external independent expertise leveraged for the identification, assessment, and evaluation of risks, and if yes, how?

⁹ See also Mapping AI Benchmark Data to Quantitative Risk Estimates (Murray et al. 2025).

¹⁰ Other examples of risk registers and incident databases include: AI Incident Database (AIID n.d.), ATLAS AI Incidents (MITRE n.d.a), MIT AI Incident Tracker (MIT 2025a), MIT AI Risk Repository (MIT 2025b), AI Risk Database (MITRE n.d.b), and AI Vulnerability Database (AVID n.d.).

TRANSPARENCY, DOCUMENTATION, AND REPORTING RECOMMENDATIONS
FOR GENERAL-PURPOSE AI RISK MANAGEMENT

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.

Measure 1.1 Mapped Recommendations

NIST AI RMF Playbook Transparency and Documentation	EU GPAI CoP (Safety and Security Chapter)	G7 (HAIP) Transparency Report Questions
<ul style="list-style-type: none"> • Did your organization address usability problems and test whether user interfaces served their intended purposes? • 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)? 	<p>for classified materials. Further, the design of the model evaluations will be informed by the model-independent information gathered.</p> <p>Measure 3.3: Signatories will conduct systemic risk modelling for the systemic risk. To this end, Signatories will:</p> <ul style="list-style-type: none"> • Use at least state-of-the-art risk modelling methods; • Build on the systemic risk scenarios developed; and • Take into account at least the information gathered [during the systemic risk identification process] and this Commitment. <p>Measure 7.3: Signatories will provide in the Model Report:</p> <ul style="list-style-type: none"> • A description of the results of their systemic risk modelling for the systemic risks; • A description of the systemic risks stemming from the model and a justification therefor, including: (i) the systemic risk estimates (pursuant to Measure 3.4); and (ii) a comparison between systemic risks with safety and security mitigations implemented and with the model fully elicited; and • All results of model evaluations relevant to understanding the systemic risks stemming from the model and descriptions of: <ul style="list-style-type: none"> ◦ How the evaluations were conducted; ◦ The tests and tasks involved in the model evaluations; ◦ How the model evaluations were scored; ◦ How the model was elicited; ◦ How the scores compare to human baselines (where applicable), across the model versions, and across the evaluation settings; ◦ A description of the access and other resources provided to: (i) internal model evaluation teams; and (ii) independent external evaluators. <p>Measure 7.4: Signatories will provide in the Model Report:</p> <ul style="list-style-type: none"> • Any available reports (e.g., via valid hyperlinks) from: (a) independent external evaluators involved in model evaluations; and • Security reviews undertaken by an independent external party. 	<ul style="list-style-type: none"> • How does your organization collaborate with relevant stakeholders across sectors to assess and adopt risk mitigation measures to address risks, in particular systemic risks? <p>Section 2:</p> <ul style="list-style-type: none"> • When does testing take place in secure environments, if at all, and if it does, how?

TRANSPARENCY, DOCUMENTATION, AND REPORTING RECOMMENDATIONS FOR GENERAL-PURPOSE AI RISK MANAGEMENT

Measure 1.1 Additional Recommendations

- Establish detailed policies and procedures that outline model evaluations processes for each risk type:
 - Evaluate models both pre- and post-mitigation (Bowen et al 2025) to account for mitigation/control failure or circumvention. This is particularly important for models intended to be released with open weights, considering the added risk of jailbreaking or modifications that may negatively affect controls in place.
 - Use closed, lighter-touch forms of testing (e.g., benchmarks) as a step prior to red teaming (TFS 2025), where certain scores trigger the need for more in-depth red teaming (Barrett et al. 2024).
 - » Document and track use of benchmarks, maintain records of erosion history, and periodically evaluate them to identify any benchmarks that may have become depreciated, outdated, or compromised.
- For each model evaluation, document and retain detailed information on:
 - Model evaluation results:
 - » Consider recording model evaluation results using evaluation disclosure cards (EvalCards) (Dhar et al. 2025), which are proposed to contain information on:
 - Modalities evaluated;
 - Languages evaluated;
 - Capability evaluations;
 - Safety evaluations; and
 - Developer footnotes.
 - » In addition to the elements included in EvalCards (Dhar et al. 2025), include sufficient information on independent third-party evaluations, including enough methodological detail to allow for independent evaluation of the results, with the exception of hazardous information or datasets that may undermine the integrity of the evaluation (Stosz et al. 2025).¹¹
 - Model evaluation methodology, including:
 - » A description of the tests and methods used, as well as justification for why they were chosen;
 - » The tasks involved in the model evaluations;
 - » An evaluation timeline that includes phases and durations for each evaluation;
 - » Limitations and challenges; and
 - » The logic behind mapping results to risk levels and risk tolerances.
 - The extent of access given to internal model evaluations teams.
 - The extent of access given to external evaluation teams.
 - Justification for moving beyond the development phase.
- In model evaluation reports (i.e., model cards and system cards), for each identified risk type, include descriptions of:
 - The evaluation methodologies;
 - Tasks involved in the evaluations;
 - Evaluation results;
 - The logic behind mapping results to risk levels and risk tolerances;
 - Limitations and challenges;
 - Justification for moving beyond the development phase;
 - Summary of third-party roles and their participation across evaluation phases;
 - Any gaps in model understanding within model documentation with clear distinctions between verified model behaviors and areas of uncertainty; and
 - The scope and limitations of evaluation coverage.
- Retain unredacted internal documentation for the purpose of sharing with oversight bodies or stakeholders with established private partnerships (e.g., third-party evaluators).

¹¹ When redacting datasets, proxy data or representative examples should be released instead (Stosz et al. 2025).

TRANSPARENCY, DOCUMENTATION, AND REPORTING RECOMMENDATIONS
FOR GENERAL-PURPOSE AI RISK MANAGEMENT

Measure 1.1 Additional Recommendations
<ul style="list-style-type: none"> Adjustments and redactions may be made depending on the nature of the stakeholder, with consideration for legal and regulatory requirements. <ul style="list-style-type: none"> In cases where redactions are necessary, provide justification and retain unredacted documentation for an appropriate period of time (e.g., five years) (California Legislature 2025, New York State Senate 2026). Redactions may be appropriate where required by state or federal law, or where necessary to protect trade secrets, cybersecurity, public safety, or U.S. national security (California Legislature 2025, New York State Senate 2026).

Risk Tracking

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.

Measure 3.2 Mapped Recommendations

NIST AI RMF Playbook Transparency and Documentation	EU GPAI CoP (Safety and Security Chapter)	G7 (HAIP) Transparency Report Questions
<p>Organizations can document the following:</p> <ul style="list-style-type: none"> Who is ultimately responsible for the decisions of the AI and is this person aware of the intended uses and limitations of the analytic? Who will be responsible for maintaining, re-verifying, monitoring, and updating this AI once deployed? To what extent does the entity communicate its AI strategic goals and objectives to the community of stakeholders? Given the purpose of this AI, what is an appropriate interval for checking whether it is still accurate, unbiased, explainable, etc.? What are the checks for this model? 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? 	<p>Measure 3.2: <i>Signatories will ensure that such model evaluations are designed and conducted using methods that are appropriate for the model and the systemic risk, and include open-ended testing of the model to improve the understanding of the systemic risk, with a view to identifying unexpected behaviours, capability boundaries, or emergent properties.</i></p> <p><i>Examples of model evaluation methods are: Q&A sets, task-based evaluations, benchmarks, red-teaming and other methods of adversarial testing, human uplift studies, model organisms, simulations, and/or proxy evaluations for classified materials. Further, the design of the model evaluations will be informed by the model-independent information gathered.</i></p> <p>Measure 3.5: <i>The following are examples of post-market monitoring methods:</i></p> <ul style="list-style-type: none"> Collecting end-user feedback; Providing (anonymous) reporting channels; Providing (serious) incident reporting forms; Providing bug bounties. 	<p>Section 2:</p> <ul style="list-style-type: none"> How do testing measures inform actions to address identified risks?

TRANSPARENCY, DOCUMENTATION, AND REPORTING RECOMMENDATIONS
FOR GENERAL-PURPOSE AI RISK MANAGEMENT

Measure 3.2 Additional Recommendations

- Establish policies and provide incentive programs for responsible disclosure of risks, vulnerabilities, and misaligned model behavior to facilitate the identification and reporting of emerging risks. Such programs could include:
 - Vulnerability disclosure programs (e.g., OpenAI 2025b);
 - Bug bounty programs:
 - » Provide bug bounty programs for high-risk use cases, e.g., Anthropic’s Agent bio bug bounty program (Anthropic 2025c);
 - Misalignment bounty programs (e.g., Palisade Research n.d.); and
 - Whistleblower protection policies (e.g., OpenAI 2024).
- Use risk registers¹² (e.g., MIT 2025b) to track and report identified risks (including difficult-to-assess risks).

Go/No-Go Decision

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.

Manage 1.1 Mapped Recommendations

NIST AI RMF Playbook Transparency and Documentation	EU GPAI CoP (Safety and Security Chapter)	G7 (HAIP) Transparency Report Questions
<p>Organizations can document the following:</p> <ul style="list-style-type: none"> • How do the technical specifications and requirements align with the AI system’s goals and objectives? • To what extent are the metrics consistent with system goals, objectives, and constraints, including ethical and compliance considerations? • What goals and objectives does the entity expect to achieve by designing, developing, and/or deploying the AI system? 	<p>Measure 4.2: Signatories will only proceed with the development, the making available on the market, and/or the use of the model, if the systemic risks stemming from the model are determined to be acceptable.</p> <p>If the systemic risks stemming from the model are not determined to be acceptable [or may soon become unacceptable], Signatories will take appropriate measures to ensure the systemic risks stemming from the model are and will remain acceptable prior to proceeding. In particular, Signatories will:</p> <ul style="list-style-type: none"> • Not make the model available on the market, restrict the making available on the market (e.g. via adjusting licenses or usage restrictions), withdraw, or recall the model, as necessary; • Implement safety and/or security mitigations (pursuant to Commitments 5 and 6); and • Conduct another round of systemic risk identification, systemic risk analysis, and systemic risk acceptance determination. 	<p>Section 2:</p> <ul style="list-style-type: none"> • How do testing measures inform actions to address identified risks?

¹² Other examples of risk registers and incident databases include the AI Incident Database (AIID n.d.), ATLAS AI Incidents (MITRE n.d.a), MIT AI Incident Tracker (MIT 2025a), MIT AI Risk Repository (MIT 2025b), AI Risk Database (MITRE n.d.b), and AI Vulnerability Database (AVID n.d.).

TRANSPARENCY, DOCUMENTATION, AND REPORTING RECOMMENDATIONS
FOR GENERAL-PURPOSE AI RISK MANAGEMENT

Manage 1.1 Additional Recommendations
<ul style="list-style-type: none"> • Based on established risk tolerances (Map 1.5) and model evaluations (Measure 1.1), determine if development and/or deployment should proceed, and document the following in detail: <ul style="list-style-type: none"> ◦ If risk is determined to be acceptable, document: <ul style="list-style-type: none"> » Justification for why risks were determined to be acceptable; » The measures in place to ensure risks remain acceptable, such as security mitigations; and » Conditions that would render the risk unacceptable. ◦ If risks are determined to be unacceptable, document: <ul style="list-style-type: none"> » Reasons the risks are unacceptable; » Planned changes to the model to bring risks down to an acceptable level; and » Plans and procedures for conducting another round of analysis and risk determination. • If risks are determined to be acceptable, provide public documentation regarding risk acceptance criteria, justification, and mitigation plans. (For more on model reports, see Govern 4.2.) • If risks are determined to be unacceptable, report results to relevant stakeholders, including oversight bodies and stakeholders with established private partnerships (e.g., third-party evaluators).

Risk Responses

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.

Manage 1.3 Mapped Recommendations

NIST AI RMF Playbook Transparency and Documentation	EU GPAI CoP (Safety and Security Chapter)	G7 (HAIP) Transparency Report Questions
<p>Organizations can document the following:</p> <ul style="list-style-type: none"> • Has the system been reviewed to ensure the AI system complies with relevant laws, regulations, standards, and guidance? • To what extent has the entity defined and documented the regulatory environment—including minimum requirements in laws and regulations? • Did your organization implement a risk management system to address risks involved in deploying the identified AI solution (e.g., personnel risk or changes to commercial objectives)? 	<p>Measure 1.1 Signatories will create a state-of-the-art Framework, taking into account the models they are developing, making available on the market, and/or using [...]. The Framework will contain:</p> <ul style="list-style-type: none"> • A high-level description of what safety and security mitigations Signatories would need to implement once each systemic risk tier is reached. <p>Measure 5.1 Signatories will implement safety mitigations that are appropriate, including sufficiently robust under adversarial pressure (e.g., fine-tuning attacks or jailbreaking), taking into account the model’s release and distribution strategy.</p> <p>Measure 6.2 Signatories will implement appropriate security mitigations to meet the Security Goal [...]. If Signatories deviate from any of the security mitigations listed [...], they will implement alternative security mitigations that achieve the respective mitigation objectives.</p>	<p>Section 2:</p> <ul style="list-style-type: none"> • What steps does your organization take to address risks and vulnerabilities across the AI lifecycle? • How does your organization promote data quality and mitigate risks of harmful bias, including in training and data collection processes? • How does your organization protect intellectual property, including copyright-protected content? • How does your organization protect privacy? How does your organization guard against systems divulging confidential or sensitive data? • How does your organization implement AI-specific information security practices pertaining to operational and cyber/physical security? • How does your organization address vulnerabilities, incidents, emerging risks? • How do testing measures inform actions to address identified risks?

TRANSPARENCY, DOCUMENTATION, AND REPORTING RECOMMENDATIONS
FOR GENERAL-PURPOSE AI RISK MANAGEMENT

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.

Manage 1.3 Mapped Recommendations

NIST AI RMF Playbook Transparency and Documentation	EU GPAI CoP (Safety and Security Chapter)	G7 (HAIP) Transparency Report Questions
	<p>Measure 7.3 Signatories will provide in the Model Report:</p> <ul style="list-style-type: none"> • A description of the systemic risks stemming from the model and a justification therefor, including: [...] a comparison between systemic risks with safety and security mitigations implemented and with the model fully elicited; • A description of: (a) all safety mitigations implemented (pursuant to Commitment 5); (b) how they fulfil the requirements of Measure 5.1; and (c) their limitations (e.g., if training on examples of undesirable model behaviour makes identifying future instances of such behaviour more difficult); and • A description of: (a) the Security Goal [...]; (b) all security mitigations implemented [...]; (c) how the mitigations meet the Security Goal, including the extent to which they align with relevant international standards or other relevant guidance [...]; and (d) if Signatories have deviated from a listed security mitigation [...], a justification for how the alternative security mitigations they have implemented achieve the respective mitigation objectives. <p>Measure 7.5 Signatories will ensure that the Model Report contains information relevant for the AI Office to understand whether and how the development, making available on the market, and/or use of the model result in material changes in the systemic risk landscape that are relevant for the implementation of systemic risk assessment and mitigation measures and processes under this Chapter. Examples of such information are:</p> <ul style="list-style-type: none"> • A description of information relevant to assessing the effectiveness of mitigations, e.g., if the model's chain-of-thought is less legible by humans. <p>Measure 10.1 Signatories will draw up and keep up-to-date the following information for the purpose of providing it to the AI Office upon request:</p> <ul style="list-style-type: none"> • A detailed description of the safety mitigations implemented. <p>Documentation will be retained at least 10 years after the model has been placed on the market. Further, Signatories will keep track of the following information, to the extent it is not already covered by the first paragraph, for the purpose of evidencing adherence to this Chapter to the AI Office upon request:</p> <ul style="list-style-type: none"> • Their processes, measures, and key decisions that form part of their systemic risk assessment and mitigation. 	<p>Section 5:</p> <ul style="list-style-type: none"> • What mechanisms, if any, does your organization put in place to allow users, where possible and appropriate, to know when they are interacting with an advanced AI system developed by your organization? • Does your organization use content provenance detection, labeling or watermarking mechanisms that enable users to identify content generated by advanced AI systems? If yes, how? Does your organization use international technical standards or best practices when developing or implementing content provenance?

TRANSPARENCY, DOCUMENTATION, AND REPORTING RECOMMENDATIONS FOR GENERAL-PURPOSE AI RISK MANAGEMENT

Manage 1.3 Additional Recommendations

- Document and retain detailed information regarding implemented safety and security mitigations, including:
 - Training data filtering, clearing, and management:
 - » How the data was evaluated and treated for bias, reliability, and representation; and
 - » If used, the safety and security practices implemented for synthetic training data.
 - Model input and output monitoring and management;
 - Model behavior management (e.g., fine tuning) to mitigate harmful responses or actions;
 - Model access management and limitation (e.g., via API);
 - Phased model release strategies;
 - Mitigation resources and tools offered to stakeholders;
 - Content/data provenance methods and techniques;
 - Interpretability or explainability methods and how they are utilized to advance safety and security;
 - Methods to prevent unauthorized access;
 - Social engineering identifiers and filters;
 - Policies and controls to protect against malware;
 - Policies and controls to protect against vulnerability explorations (e.g., regular software updates and patch management);
 - Mitigations implemented to protect unreleased model parameters (see General-Purpose AI Code of Practice, Safety and Security Chapter, Appendix 4.2 and 4.3, EC 2025a).
 - Protections against insider threats (see General-Purpose AI Code of Practice, Safety and Security Chapter, Appendix 4.4, EC 2025a).
- In model reports (i.e., model cards and system cards), include descriptions of:
 - Implemented security mitigations and controls:
 - » When publicly reporting security mitigations and controls, do not provide information that may potentially benefit threat actors.
 - Assessments and adequacy of the implemented mitigations and controls (California Legislature 2025);
 - The use of third-parties to evaluate model risk (California Legislature 2025); and
 - Implemented practices to prevent unauthorized access, modification, or transfer of secure unreleased model weights by internal and external parties (California Legislature 2025).
- Retain internal documentation for the purpose of sharing with oversight bodies or stakeholders with established private partnerships (e.g., third-party evaluators).
- Partner with relevant stakeholders to share research and best practices on risk mitigations.
- Adjustments and redactions may be made depending on the nature of the stakeholder, with consideration for legal and regulatory requirements.
 - In cases where redactions are necessary, provide justification and retain unredacted documentation for an appropriate period of time (e.g., five years) (California Legislature 2025).
- Share with appropriate oversight offices risk evaluation methods and results for internal use of the model, including risks related to the model circumventing oversight mechanisms (California Legislature 2025).
 - Share the results pursuant of a pre-defined cadence (e.g., three months).

TRANSPARENCY, DOCUMENTATION, AND REPORTING RECOMMENDATIONS
FOR GENERAL-PURPOSE AI RISK MANAGEMENT

Responses to Emergent Risk

Manage 2.3: Procedures are followed to respond to and recover from a previously unknown risk when it is identified.		
Manage 2.3 Mapped Recommendations		
NIST AI RMF Playbook Transparency and Documentation	EU GPAI CoP (Safety and Security Chapter)	G7 (HAIP) Transparency Report Questions
<p>Organizations can document the following:</p> <ul style="list-style-type: none"> • Who will be responsible for maintaining, re-verifying, monitoring, and updating this AI once deployed? • Are the responsibilities of the personnel involved in the various AI governance processes clearly defined? (Including responsibilities to decommission the AI system.) • What processes exist for data generation, acquisition/ collection, ingestion, staging/ storage, transformations, security, maintenance, and dissemination? • How will the appropriate performance metrics, such as accuracy of the AI, be monitored after the AI is deployed? 	<p>Measure 1.3 Signatories will conduct an appropriate Framework assessment, if they have reasonable grounds to believe that the adequacy of their Framework and/or their adherence thereto has been or will be materially undermined [...]. Examples of such grounds are:</p> <ul style="list-style-type: none"> • The systemic risks stemming from at least one of their models have changed or are likely to change materially, e.g., safety and/ or security mitigations have become or are likely to become materially less effective, or at least one of their models has developed or is likely to develop materially changed capabilities and/or propensities. <p>Measure 7.6 Signatories will update their Model Report if they have reasonable grounds to believe that the justification for why the systemic risks stemming from the model are acceptable [...] has been materially undermined. Examples of such grounds are:</p> <ul style="list-style-type: none"> • The model's capabilities, propensities, and/or affordances have changed or will change materially, such as through further post-training, access to additional tools, or increase in inference compute. 	<p>Section 2:</p> <ul style="list-style-type: none"> • What steps does your organization take to address risks and vulnerabilities across the AI lifecycle?

Manage 2.3 Additional Recommendations
<ul style="list-style-type: none"> • Develop and document practices and procedures for continuous monitoring and rapid response to unknown risk: <ul style="list-style-type: none"> ◦ Monitor risk registers and incident databases. ◦ Continuously monitor the model using the methods outlined in Manage 4.1 (Post-Deployment Monitoring). <ul style="list-style-type: none"> » To determine safety measures function properly and are capable of blocking harms, post-deployment evaluation and monitoring are needed to protect more capable models. Reporting can be shared with the public without disclosing precise performance metrics, or with trusted government entities on specific methods and results (Bowen et al 2025). ◦ Document any new risks and retain documentation to track risk progression. ◦ Report new risks, and related incidents, to public repositories and databases to promote information sharing, prevent reporting gaps, and enhance knowledge around harms and risks (Dixon and Frase 2025).¹³ ◦ If the newly discovered risk surpasses pre-established risk tolerance, halt development, restrict use, or take the model offline if necessary while risk responses are being evaluated.

¹³ Examples of risk registers and incident databases include the AI Incident Database (AIID n.d.), ATLAS AI Incidents (MITRE n.d.a), MIT AI Incident Tracker (MIT 2025a), MIT AI Risk Repository (MIT 2025b), AI Risk Database (MITRE n.d.b), and AI Vulnerability Database (AVID n.d.).

TRANSPARENCY, DOCUMENTATION, AND REPORTING RECOMMENDATIONS FOR GENERAL-PURPOSE AI RISK MANAGEMENT

Manage 2.3 Additional Recommendations

- Re-evaluate the risk management framework and update if necessary.
 - » For example, Anthropic’s Responsible Scaling Policy v2.2 establishes capability thresholds and revisits ASL-3 Required Safeguards at least annually to ensure appropriate implementation (Anthropic 2025b).¹⁴
- Update public-facing documentation (e.g., model/system cards and safety frameworks) to reflect discovered risks.
 - » For example, Google updated its Frontier Safety Framework to add new risk domains (e.g., harmful manipulation, misalignment) and refine its risk assessment process (Flynn et al 2025).
- Communicate any previously unknown risks, sources of the risk, and planned mitigations to relevant stakeholders (e.g., model developers, oversight bodies).

Decommissioning Mechanisms

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.

Manage 2.4 Mapped Recommendations

NIST AI RMF Playbook Transparency and Documentation	EU GPAI CoP (Safety and Security Chapter)	G7 (HAIP) Transparency Report Questions
<p>Organizations can document the following:</p> <ul style="list-style-type: none"> • What are the roles, responsibilities, and delegation of authorities of personnel involved in the design, development, deployment, assessment, and monitoring of the AI system? • Did your organization implement a risk management system to address risks involved in deploying the identified AI solution (e.g., personnel risk or changes to commercial objectives)? • 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 does the entity have established procedures for retiring the AI system, if it is no longer needed? • How did the entity use assessments and/or evaluations to determine if the system can be scaled up, continue, or be decommissioned? 	<p>Measure 4.2</p> <p>Signatories will only proceed with the development, the making available on the market, and/or the use of the model, if the systemic risks stemming from the model are determined to be acceptable.</p> <p>If the systemic risks stemming from the model are not determined to be acceptable or are reasonably foreseeable to be soon not determined to be acceptable, Signatories will take appropriate measures to ensure the systemic risks stemming from the model are and will remain acceptable prior to proceeding. In particular, Signatories will:</p> <ul style="list-style-type: none"> • Not make the model available on the market, restrict the making available on the market (e.g., via adjusting licenses or usage restrictions), withdraw, or recall the model, as necessary; • Implement safety and/or security mitigations • Conduct another round of systemic risk identification, systemic risk analysis, and systemic risk acceptance determination. 	<p>Section 2:</p> <ul style="list-style-type: none"> • How does your organization address vulnerabilities, incidents, emerging risks?

¹⁴ The most recent update of Anthropic’s Responsible Scaling Policy however does not include this provision (Anthropic 2026).

TRANSPARENCY, DOCUMENTATION, AND REPORTING RECOMMENDATIONS
FOR GENERAL-PURPOSE AI RISK MANAGEMENT

Manage 2.4 Additional Recommendations

- Establish policies and procedures to halt development, halt deployment, or recall the model if risks are determined to be unacceptable after the appropriate mitigations, controls, and adjustments have been implemented. These policies should include:
 - Clearly defined roles, responsibilities, and lines of communication;
 - Training requirements for personnel and technical teams;
 - Requirements for back-up systems and required infrastructure;
 - Criteria and mechanisms for automatic shutoff; and
 - Mechanisms for manual overrides.
- In Terms of Use policies and other appropriate public documentation, include terms that address the possibility of a deployed model being recalled:
 - Inform users how a recall would affect user data, subscriptions, and access to any linked services and applications.
 - Notify users when a deployed model is scheduled to be recalled or suspended.
- Report model recalls, halts to development, and halts to deployment to relevant oversight bodies and affected stakeholders.

Post-Deployment Monitoring

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.*

Manage 4.1 Mapped Recommendations

NIST AI RMF Playbook Transparency and Documentation	EU GPAI CoP (Safety and Security Chapter)	G7 (HAIP) Transparency Report Questions
<p><i>Organizations can document the following:</i></p> <ul style="list-style-type: none"> • <i>To what extent has the entity documented the post-deployment AI system’s testing methodology, metrics, and performance outcomes?</i> • <i>How easily accessible and current is the information available to external stakeholders?</i> 	<p>Measure 3.5</p> <p><i>Signatories will conduct appropriate post-market monitoring to gather [and document] information relevant to assessing whether the systemic risk could be determined to not be acceptable and to inform whether a Model Report update is necessary.</i></p> <p><i>To these ends, post-market monitoring will:</i></p> <ul style="list-style-type: none"> • <i>Gather information about the model’s capabilities, propensities, affordances, and/or effects;</i> • <i>Take into account the exemplary methods listed below; and</i> • <i>If Signatories themselves provide and/or deploy AI systems that integrate their own model, include monitoring the model as part of these AI systems.</i> <p><i>The following are examples of post-market monitoring methods [...]:</i></p> <ul style="list-style-type: none"> • <i>Collecting end-user feedback;</i> • <i>Providing (anonymous) reporting channels;</i> • <i>Providing (serious) incident reporting forms;</i> • <i>Providing bug bounties;</i> • <i>Establishing community-driven model evaluations and public leaderboards;</i> 	<p>Section 1:</p> <ul style="list-style-type: none"> • <i>What practices does your organization use to identify and evaluate risks such as vulnerabilities, incidents, emerging risks and misuse, throughout the AI lifecycle?</i>

TRANSPARENCY, DOCUMENTATION, AND REPORTING RECOMMENDATIONS
FOR GENERAL-PURPOSE AI RISK MANAGEMENT

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.*

Manage 4.1 Mapped Recommendations

NIST AI RMF Playbook Transparency and Documentation	EU GPAI CoP (Safety and Security Chapter)	G7 (HAIP) Transparency Report Questions
	<ul style="list-style-type: none"> • <i>Conducting frequent dialogues with affected stakeholders;</i> • <i>Monitoring software repositories, known malware, public forums, and/or social media for patterns of use;</i> • <i>Supporting the scientific study of the model’s capabilities, propensities, affordances, and/or effects in collaboration with academia, civil society, regulators, and/or independent researchers;</i> • <i>Implementing privacy-preserving logging and metadata analysis techniques of the model’s inputs and outputs using, e.g., watermarks, metadata, and/or other at least state-of-the-art provenance techniques;</i> • <i>Collecting relevant information about breaches of the model’s use restrictions and subsequent incidents arising from such breaches; and/or</i> • <i>Monitoring aspects of models that are relevant for assessing and mitigating systemic risk and are not transparent to third parties, e.g., hidden chains-of-thought for models for which the parameters are not publicly available for download.</i> <p><i>To facilitate post-market monitoring, Signatories will provide an adequate number of independent external evaluators with adequate free access to:</i></p> <ul style="list-style-type: none"> • <i>The model’s most capable model version(s) with regard to the systemic risk that is made available on the market;</i> • <i>The chains-of-thought of the model version(s) in point (1), if available; and</i> • <i>The model version(s) corresponding to the [most capable] model version(s) [...] with the fewest safety mitigations implemented with regard to the systemic risk (such as the helpful-only model version, if it exists) and, as available, its chains-of-thought.</i> 	

TRANSPARENCY, DOCUMENTATION, AND REPORTING RECOMMENDATIONS
FOR GENERAL-PURPOSE AI RISK MANAGEMENT

Manage 4.1 Additional Recommendations

- Establish policies and procedures for post-deployment monitoring to gather relevant information for continued risk assessment of the model.
 - Collect user feedback by providing comprehensive and easily accessible third-party reporting mechanisms that include:
 - » Reporting forms for serious incidents;
 - » Anonymous reporting channels; and
 - » Vulnerability disclosure and bug bounty programs (see, Measure 3.2).
 - Implement bi-directional feedback mechanisms that facilitate active engagement and an iterative exchange of information.
- Report any incidents, near misses, discovered risks, and vulnerabilities to public repositories and databases¹⁵ (e.g., AIID n.d., MIT 2025b).
 - When reporting incidents, include the following information:
 - » Date of the incident (California Legislature 2025, New York State Senate 2026);
 - » A description of the incident (California Legislature 2025, New York State Senate 2026);
 - » The reasons the incident qualifies as a safety incident (California Legislature 2025, New York State Senate 2026); and
 - » Whether or not this was an incident associated with internal use of the model (California Legislature 2025, New York State Senate 2026).

¹⁵ Other examples of risk registers and incident databases include the AI Incident Database (AIID n.d.), ATLAS AI Incidents (MITRE n.d.a), MIT AI Incident Tracker (MIT 2025a), MIT AI Risk Repository (MIT 2025b), AI Risk Database (MITRE n.d.b), and AI Vulnerability Database (AVID n.d.).

Glossary

KEY TERMS AND DEFINITIONS

- **CoP:** Code of Practice
- **HAIP:** Hiroshima AI Process
- **Transparency:** Broad information sharing with relevant stakeholders and the public.
- **Documentation:** The process of developing structured artifacts that record and communicate technical, operational, and contextual information.
- **Reporting:** Formal sharing of specific information, often in accordance with defined templates, protocols, regulatory requirements, or contractual obligations.
- **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)).

Appendices

APPENDIX 1: UNMAPPED QUESTIONS FROM THE HAIP REPORTING FRAMEWORK

Guidance for the sections and questions from the G7 HAIP reporting framework (OECD.AI 2025) that did not directly map to any of our GPAI Profile high-priority sub-categories are listed below. While these questions and corresponding recommendations did not appropriately align, we recommended that they still be considered when establishing transparency, documentation, and reporting practices.

Unmapped questions from the HAIP Reporting Framework:

- **Section 6:** *What research or investment is your organization pursuing to minimize socio-economic and/or environmental risks from AI?*
- **Section 6:** *How does your organization advance research and investment related to the following: security, safety, bias and disinformation, fairness, explainability and interpretability, transparency, robustness, and/or trustworthiness of advanced AI systems?*
- **Section 6:** *How does your organization collaborate on and invest in research to advance the state of content authentication and provenance?*
- **Section 6:** *Does your organization participate in projects, collaborations, and investments in research that support the advancement of AI safety, security, and trustworthiness, as well as risk evaluation and mitigation tools?*
- **Section 7:** *What research or investment is your organization pursuing to maximize socio-economic and environmental benefits from AI? Please provide examples.*
- **Section 7:** *Does your organization support any 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? Please provide examples.*
- **Section 7:** *Does your organization prioritize AI projects for responsible stewardship of trustworthy and human-centric AI in support of the UN Sustainable Development Goals? Please provide examples.*

TRANSPARENCY, DOCUMENTATION, AND REPORTING RECOMMENDATIONS
FOR GENERAL-PURPOSE AI RISK MANAGEMENT

Recommendations:

Consider establishing and implementing policies for collaboration, support, and investment in research for:

- 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.

APPENDIX 2: ADDITIONAL RESOURCES

Transparency, Documentation, and Reporting Initiatives and Frameworks

Resource Name	Resource Type
EU reporting template for serious incidents involving general-purpose AI models with systemic risk (EC 2025b)	Serious Incident Report Draft Template
G7 reporting framework – Hiroshima AI Process (HAIP) international code of conduct for organizations developing advanced AI systems (G7 2023)	Model Reporting Framework
The General-Purpose AI Code of Practice (EC 2025a) Transparency chapter Model Documentation Form	Model Documentation Template
NIST Extended Outline: Proposed Zero Draft for a Standard on Documentation of AI Datasets and AI Models (NIST 2025)	Model Documentation Pre-Standard (draft outline)
OECD Towards a Common Reporting Framework for AI Incidents (OECD 2025b)	Incident Reporting Framework Criteria
Responsible AI Safety and Education Act (New York State Senate 2026)	Transparency and Reporting Law
Transparency in Frontier Artificial Intelligence Act, S.B. 53 (California Legislature 2025)	Transparency and Reporting Law

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 General-Purpose AI (GPAI) Risk-Management Standards Profile, 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 Saliia Asanova, Evelina Ayrapetyan, Kathy Baxter, Kendrea Beers, Haydn Belfield, Marta Bieńkiewicz, Marjory Blumenthal, Miranda Bogen, Sean Brooks, Siméon Campos, Ryan Carrier, Jonathan Cefalu, Colleen Chien, Ze Shen Chin, Joe Collman, Talita Dias, Drenan Dudley, Ian Eisenberg, Aryeh Englander, Alex Engler, Yoav Evenstein, Audrie Francis, Andrew Gamino-Cheong, James Gealy, Thomas Gilbert, AJ Grotto, Koen Holtman, Steph Ifayemi, Nikhil Jain, Caroline Jeanmaire, Jessica Ji, Zaheed Kara, Leonie Koessler, Noam Kolt, Viktoriia Kravchyk, Benjamin Larsen, Meredith Lee, Natalia Luka, Oumou Ly, Devin Lynch, Richard Mallah, Deirdre Mulligan, Malcolm Murray, Julia Mykhailiuk, Mina Narayanan, Elaine Newton, David Norman, Joe O'Brien, Amin Oueslati, Milan Patel, Matteo Pistillo, Anka Reuel, Stuart Russell, Krishna Sankar, Daniel Schiff, Lea Shanley, Raymond Sheh, Buck Shlegeris, Aparajita Singh, Peter Slattery, Andrew Smart, Genevieve Smith, Adriana Stephan, Zeerak Talat, Nel Talverdi, Esther Tetrushvily, Kristen Vrionis, Victor Zhenyi Wang, Kevin Wei, Laurin Weissinger, Devon Whittle, Cherry Wu, Andy Yang, and Lenora Zimmerman.

This work was financially supported by funding from Coefficient Giving.

References

- AIID (n.d.) AI Incident Database. AIID, <https://incidentdatabase.ai/>
- Anthropic (n.d.a) Research at Anthropic. Anthropic, <https://www.anthropic.com/research>
- Anthropic (n.d.b) Anthropic's Transparency Hub. Anthropic, <https://www.anthropic.com/transparency>
- Anthropic (2025a) System Card Addendum: Claude Opus 4.1. Anthropic, <https://assets.anthropic.com/m/4c024b86c698d3d4/original/Claude-4-1-System-Card.pdf>
- Anthropic (2025b) Responsible Scaling Policy. Anthropic, <https://www-cdn.anthropic.com/872c653b2d0501d6ab44cf87f43e1dc4853e4d37.pdf>
- Anthropic (2025c) Agent bio bug bounty. Anthropic, <https://openai.com/bio-bug-bounty/>
- Anthropic (2025d) Proposed Frontier Model Transparency Framework. Anthropic, <https://www-cdn.anthropic.com/19cc4bf9eb6a94f9762ac67368f3322cf82b09fe.pdf>
- Anthropic (n.d.) Anthropic Economic Index. Anthropic, <https://www.anthropic.com/economic-index>
- Anthropic (2026) Responsible Scaling Policy v3.0. Anthropic, <https://www-cdn.anthropic.com/e670587677525f28df69b59e5fb4c22cc5461a17.pdf>
- Chloe Autio, Reva Schwartz, Jesse Dunietz, Shomik Jain, Martin Stanley, Elham Tabassi, Patrick Hall, 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>
- AVID (n.d.) AI Vulnerability Database. AVID, <https://avidml.org/>
- Anthony M. Barrett, Krystal Jackson, Evan R. Murphy, Nada Madkour, 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>
- 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 Molchanovskiy, Andrea Monti, Nusu Mwamanzu, 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, Yi Zeng (2025) International AI Safety Report, <https://arxiv.org/abs/2501.17805>
- Rishi Bommasani, Kevin Klyman, Sayash Kapoor, Shayne Longpre, Betty Xiong, Nestor Maslej, Percy Liang (2024a) The 2024 Foundation Model Transparency Index. arXiv, <https://arxiv.org/abs/2407.12929>

TRANSPARENCY, DOCUMENTATION, AND REPORTING RECOMMENDATIONS FOR GENERAL-PURPOSE AI RISK MANAGEMENT

- Rishi Bommasani, Kevin Klyman, Shayne Longpre, Betty Xiong, Sayash Kapoor, Nestor Maslej, Arvind Narayanan, Percy Liang (2024b) Foundation Model Transparency Reports. arXiv, <https://arxiv.org/pdf/2402.16268>
- Dillon Bowen, Ann-Kathrin Dombrowski, Adam Gleave, Chris Cundy (2025) AI Companies Should Report Pre- and Post-Mitigation Safety Evaluations. arXiv, <https://arxiv.org/pdf/2503.17388>
- California Legislature (2025). Transparency in Frontier Artificial Intelligence Act, S.B. 53. California Legislature, https://leginfo.ca.gov/faces/billTextClient.xhtml?bill_id=202520260SB53
- Kasia Chmielinski, Sarah Newman, Chris N. Kranzinger, Michael Hind, Jennifer Wortman, Vaughan, Margaret Mitchell, Julia Stoyanovich, Angelina McMillan-Major, Emily McReynolds, Kathleen Esfahany, Mary L. Gray, Audrey Chang, Maui Hudson (2024) The CLeAR Documentation Framework for AI Transparency: Recommendations for Practitioners & Context for Policymakers. The Shorenstein Center on Media, Politics and Public Policy, <https://shorensteincenter.org/clear-documentation-framework-ai-transparency-recommendations-practitioners-context-policymakers/>
- Kyle Crichton, Abhiram Reddy, Jessica Ji, Ali Crawford, Mia Hoffmann, Colin Shea-Blymyer, John Bansemer (2025) Harmonizing AI Guidance: Distilling Voluntary Standards and Best Practices into a Unified Framework. Center for Security and Emerging Technology, <https://cset.georgetown.edu/publication/harmonizing-ai-guidance-distilling-voluntary-standards-and-best-practices-into-a-unified-framework>
- 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>
- Ruchira Dhar, Danae Sanchez Villegas, Antonia Karamolegkou, Alice Schiavone, Yifei Yuan, Xinyi Chen, Jiaang Li, Stella Frank, Laura De Grazia, Monorama Swain, Stephanie Brandl, Daniel Hershovich, Anders Sogaard, Desmond Elliott (2025) EvalCards: A Framework for Standardized Evaluation Reporting. arXiv, <https://arxiv.org/abs/2511.21695>
- Ren Bin Lee Dixon and Heather Frase (2025) AI Incidents Key Components for a Mandatory Reporting Regime. Center for Security and Emerging Technology, <https://cset.georgetown.edu/wp-content/uploads/CSSET-AI-Incidents.pdf?utm>
- EC (2025a) The General-Purpose AI Code of Practice. European Commission, <https://digital-strategy.ec.europa.eu/en/policies/contents-code-gpai>
- EC (2025b) AI Act: Commission Publishes a Reporting Template for Serious Incidents Involving General-Purpose AI Models with Systemic Risk. European Commission, <https://digital-strategy.ec.europa.eu/en/library/ai-act-commission-publishes-reporting-template-serious-incidents-involving-general-purpose-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>
- Four Flynn, Helen King, Anca Dragan (2025) Strengthening our Frontier Safety Framework. Google DeepMind, <https://deepmind.google/discover/blog/strengthening-our-frontier-safety-framework/>
- 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>
- Google (2025a) Gemini 2.5 Pro Model Card. Google, <https://storage.googleapis.com/deepmind-media/Model-Cards/Gemini-2-5-Pro-Model-Card.pdf>

TRANSPARENCY, DOCUMENTATION, AND REPORTING RECOMMENDATIONS FOR GENERAL-PURPOSE AI RISK MANAGEMENT

- Google (2025b) 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 (2025c) Responsible AI Progress Report. Google, <https://ai.google/static/documents/ai-responsibility-update-published-february-2025.pdf>
- Google DeepMind (n.d.) Research, <https://deepmind.google/research/>
- Balint Gyevnara, Nick Ferguson, Burkhard Schafer (2023) Bridging the Transparency Gap: What Can Explainable AI Learn From the AI Act? arXiv, <https://arxiv.org/pdf/2302.10766>
- John Howell and Stephanie Ifayemi (2024) Policy Alignment on AI Transparency. Partnership on AI, https://partnershiponai.org/wp-content/uploads/2024/09/PAI_Policy-Alignment.pdf
- ISO (2023) ISO/IEC 23894:2023 Information technology — Artificial intelligence — Guidance on risk management. ISO, <https://www.iso.org/standard/77304.html>
- 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>
- Shayne Longpre, Sayash Kapoor, Kevin Klyman, Ashwin Ramaswami, Rishi Bommasani, Borhane Bllili-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, 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 Bllili-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 Mialhe, Nik Marda, Peter Henderson, Rebecca S. Portnoff, Rebecca Weiss, Victoria Westerhoff, Yacine Jernite, Rumman Chowdhury, Percy Liang, 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>
- Laura Lucaj, Alex Loosley, Hakan Jonsson, Urs Gasser, Patrick van der Smagt (2025) TechOps: Technical Documentation Templates for the AI Act. arXiv, <https://arxiv.org/pdf/2508.08804>
- Nada Madkour, Jessica Newman, Deepika Raman, Krystal Jackson, Evan R. Murphy, Charlotte Yuan, Dan Hendrycks (2026) General Purpose AI Risk-Management Standards Profile, Version 1.2. UC Berkeley Center for Long-Term Cybersecurity, <https://cltc.berkeley.edu/publication/ai-risk-management-standards-profile-v1.2/>
- Rita Matulionyte (2023) Regulating Transparency of AI: A Survey of Best Practices. SSRN, https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4554868
- Meta (n.d.) Model Cards & Prompt formats: Llama 4. Meta, <https://www.llama.com/docs/model-cards-and-prompt-formats/llama4/>
- Microsoft (2025a) 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>
- Microsoft (2025b) 2025 Responsible AI Transparency Report. Microsoft, <https://cdn-dynmedia-1.microsoft.com/is/content/microsoftcorp/microsoft/msc/documents/presentations/CSR/Responsible-AI-Transparency-Report-2025-vertical.pdf>
- MIT (2025a) AI Incident Tracker. MIT <https://airisk.mit.edu/ai-incident-tracker>
- MIT (2025b) What are the risks from Artificial Intelligence? MIT, <https://airisk.mit.edu/>

TRANSPARENCY, DOCUMENTATION, AND REPORTING RECOMMENDATIONS
FOR GENERAL-PURPOSE AI RISK MANAGEMENT

- MITRE (n.d.a) MITRE ATLAS AI Incidents. MITRE, <https://ai-incidents.mitre.org/>
- MITRE (n.d.b) AI Risk Database. MITRE, <https://ai-risk.mitre.org/>
- 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>
- New York State Senate (2026) Responsible AI Safety and Education Act (RAISE Act). New York State Senate, <https://www.nysenate.gov/legislation/bills/2025/S8828>
- NIST (2023a) AI 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 (2025) Extended Outline: Proposed Zero Draft for a Standard on Documentation of AI Datasets and AI Models. National Institute of Standards and Technology, https://www.nist.gov/system/files/documents/2025/09/12/Extended%20Outline%20-%20Proposed%20Zero%20Draft%20for%20a%20Standard%20on%20Documentation%20of%20AI%20datasets%20and%20AI%20models_September%202025.pdf
- Jessica Newman, Deepika Raman, Krystal A. Jackson, Nada Madkour, Evan R. Murphy (2025) How to Say No to the Next AI Release. Tech Policy Press, <https://www.techpolicy.press/how-to-say-no-to-the-next-ai-release/>
- OECD (2022) Oecd Framework for the Classification of AI Systems. OECD Publishing, https://www.oecd.org/content/dam/oecd/en/publications/reports/2022/02/oecd-framework-for-the-classification-of-ai-systems_336a8b57/cb6d9eca-en.pdf
- OECD (2025a) How Are AI Developers Managing Risks? Insights From Responses to the Reporting Framework of The Hiroshima AI Process Code of Conduct. OECD Publishing, https://www.oecd.org/content/dam/oecd/en/publications/reports/2025/09/how-are-ai-developers-managing-risks_fbaeb3ad/658c2ad6-en.pdf
- OECD (2025b) Towards a Common Reporting Framework for AI Incidents. OECD Publishing, https://www.oecd.org/content/dam/oecd/en/publications/reports/2025/02/towards-a-common-reporting-framework-for-ai-incidents_8c488fdb/f326d4ac-en.pdf
- OECD.AI (n.d.) Submitted Reports. OECD Policy Observatory, <https://transparency.oecd.ai/reports>
- 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/>
- OpenAI (n.d.) Trust & transparency. OpenAI, <https://openai.com/trust-and-transparency/>
- OpenAI (2024) OpenAI’s Raising Concerns Policy. OpenAI, <https://openai.com/index/openai-raising-concerns-policy/>
- OpenAI (2025a) GPT-5 System Card. OpenAI, <https://openai.com/index/gpt-5-system-card/>
- OpenAI (2025b) Coordinated vulnerability disclosure policy. OpenAI, <https://openai.com/policies/coordinated-vulnerability-disclosure-policy/>
- OpenAI (2025c) Preparedness Framework. OpenAI, <https://cdn.openai.com/pdf/18ao2b5d-6b67-4cec-ab64-68cdfbddebcd/preparedness-framework-v2.pdf>
- Palisade Research (n.d.) AI Misalignment Bounty. Palisade Research, <https://bounty.palisaderesearch.org/>
- Irene Solaiman (2023) The Gradient of Generative AI Release: Methods and Considerations. arXiv, <https://arxiv.org/pdf/2302.04844>
- Conrad Stosz, Karson Elmgren, Charles Foster, George Balston, Seth Donoughe, Samira Nedungadi, Michael Chen, Jasper Götting, Patricia Paskov, Sayash Kapoor, Sarah Schwettmann, Rishi Bommasani, Luca Righetti, Sean McGregor, Grace Werner, Rob Reich, Arvind Narayanan, Elizabeth Barnes, Christopher Painter, Miles

TRANSPARENCY, DOCUMENTATION, AND REPORTING RECOMMENDATIONS
FOR GENERAL-PURPOSE AI RISK MANAGEMENT

- Brundage, Aidan Homewood, Divya Siddharth, Faisal Lalani, Charles Teague, Jaime Sevilla, Jacob Steinhardt (2025) AEF-1: Minimum Operating Conditions for Independent Third Party AI Evaluations. AEF-1 Standard, <https://www.aef.one/aef-one.pdf>
- TFS (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>
- Alexander Wan, Kevin Klyman, Sayash Kapoor, Nestor Maslej, Shayne Longpre, Betty Xiong, Percy Liang, Rishi Bommasani (2025) The 2025 Foundation Model Transparency Index. arXiv, <https://arxiv.org/abs/2512.10169>
- Jennifer Wang, Kayla Huang, Kevin Klyman, Rishi Bommasani (2025) Do AI Companies Make Good on Voluntary Commitments to the White House? Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society, <https://ojs.aaai.org/index.php/AIES/article/view/36743>
- Amy A. Winecoff and Miranda Bogen (2024) Improving governance outcomes through AI documentation: Bridging theory and practice. arXiv, <https://arxiv.org/abs/2409.08960>
- Anna Katariina Wisakanto, Joe Rogero, Avyay M. Casheekar, Richard Mallah (2025) Adapting Probabilistic Risk Assessment for AI. arXiv, <https://arxiv.org/abs/2504.18536>



CLTC

Center for Long-Term
Cybersecurity

UC Berkeley