UC BERKELEY

CENTER FOR LONG-TERM CYBERSECURITY

## QUICK GUIDE

# An Introductory Resource for the AI Risk-Management Standards Profile for General-Purpose AI (GPAI) and Foundation Models

ANTHONY M. BARRETT | JESSICA NEWMAN | BRANDIE NONNECKE | NADA MADKOUR
DAN HENDRYCKS | EVAN R. MURPHY | KRYSTAL JACKSON | DEEPIKA RAMAN

**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 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.

# QUICK GUIDE

# An Introductory Resource for the AI Risk-Management Standards Profile for General-Purpose AI (GPAI) and Foundation Models

**ANTHONY M. BARRETT**[†] • **JESSICA NEWMAN**[†] • **BRANDIE NONNECKE**[††] • **NADA MADKOUR**[†] • **DAN HENDRYCKS**[†††] • **EVAN R. MURPHY**[†] • **KRYSTAL JACKSON**[†] • **DEEPIKA RAMAN**[†]

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

†† CITRIS Policy Lab, CITRIS and the Banatao Institute; Goldman School of Public Policy, UC Berkeley

††† Berkeley AI Research Lab, UC Berkeley

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

**The Quick Guide is a short and actionable introductory resource designed to complement the full Profile. For the full AI Risk-Management Standards Profile for General-Purpose AI (GPAI) and Foundation Models V1.1, see:**

https://cltc.berkeley.edu/publication/ai-risk-management-standards-profile-v1-1

CLTC Center for Long-Term Cybersecurity UC Berkeley

AI SECURITY INITIATIVE

CITRIS AND THE BANATAO INSTITUTE | CITRIS POLICY LAB

# Contents

# 1. How to Use this Quick Guide

This quick guide is designed to be a short and actionable introductory resource to complement the full AI Risk-Management Standards Profile for General-Purpose AI (GPAI) and Foundation Models (Profile) V1.1, available here: https://cltc.berkeley.edu/wp-content/uploads/2025/01/Berkeley-AI-Risk-Management-Standards-Profile-for-General-Purpose-AI-and-Foundation-Models-v1-1.pdf.

**Please refer to the full Profile for sources, resources, and significantly more detail on all of the recommendations listed here, as well as other related guidance.**

Both the Quick Guide and Profile are intended to be used in conjunction with the NIST AI Risk Management Framework (AI RMF) and AI RMF Playbook or an approximately equivalent set of AI risk management guidance resources.

We provide guidance specifically for general-purpose AI (GPAI) and foundation models, across the four AI RMF "core functions," including: Govern, Map, Measure, and Manage.

We expect this guide to be primarily helpful for developers of large-scale, state-of-the-art GPAI/foundation models. Others who can benefit from use of this guide include downstream developers of end-use applications or AI systems that build on a GPAI/foundation model, as well as model evaluators and regulators.

## 1.1 WHAT'S INCLUDED

The quick guide was developed to be a condensed version of the full Profile, and includes:

- **Highest-priority risk management steps** to be regarded as baseline expectations. These are the same as those outlined in the full Profile. Further detail, including inclusion criteria for high-priority risk management steps, can be found in section 2.3 of the full V1.1 Profile.
- A **"Topic Table"** that maps various topics to sub-sections of the core RMF functions.
- **Priority guidance in condensed form** for each of the core RMF functions, with "high-priority risk management steps" highlighted for ease of reference.

## 1.2 RECOMMENDED INSTRUCTIONS

- Read section 2, "High-Priority Risk Management Steps"
- For specific topics, use the Topic Table in section 3 to identify the sub-categories relevant to your chosen topic(s) and core function(s). For general application of the Profile, the Topic Table may be skipped.
- Read the shortened high-level guidance in section 4. You may take note of particularly relevant subsections to refer to when reading the full Profile guidance, or go through the full Profile in parallel to reading section 4 and identifying particularly relevant subsections.
  - » High-priority subsections[1] are highlighted; we encourage you to prioritize them while applying the quick guide and the full Profile.

***The quick guide is meant to introduce readers to the Profile V1.1 contents and aid in Profile V1.1 navigation. We do not recommend using the quick guide as a replacement for the full Profile.***

1 All the subsections considered "high-priority" in the quick guide are considered high priority in the full Profile.

# 2. High-Priority Risk Management Steps

We recommend giving the highest priority to the following risk management steps and corresponding Profile guidance sections.[2] (Appropriately applying the Profile guidance for the following steps should be regarded as the baseline or minimum expectations for users of this Profile, who can exceed the minimum expectations by also applying guidance in other sections.)

- **Check or update, and incorporate, each of the high-priority risk management steps in this list when making go/no-go decisions,** especially on whether to proceed on major stages or investments for development or deployment of cutting-edge large-scale GPAI/ foundation models (Manage 1.1).
- **Take responsibility for risk assessment and risk management tasks for which your organization has access to information, capability, or opportunity to develop capability sufficient for constructive action, or that is substantially greater than others in the value chain** (Govern 2.1).
  - » We also recommend applying this principle throughout other risk assessment and risk management steps, and we refer to it frequently in other guidance sections.
- **Set risk-tolerance thresholds to prevent unacceptable risks** (Map 1.5).
  - » For example, the NIST AI RMF 1.0 recommends the following: "In cases where an AI system presents unacceptable negative risk levels – such as where significant negative impacts are imminent, severe harms are actually occurring, or catastrophic risks are present – development and deployment should cease in a safe manner until risks can be sufficiently managed" (NIST 2023a, p.8).
- **Identify reasonably foreseeable uses, misuses, and abuses for a GPAI/foundation model** (e.g., automated generation of toxic or illegal content or disinformation, or aiding with proliferation of cyber, chemical, biological, or radiological weapons), and identify reasonably foreseeable potential impacts (e.g., to fundamental rights) (Map 1.1).
- **Identify whether a GPAI/foundation model could lead to significant, severe, or catastrophic impacts,** e.g., because of correlated failures or errors across high-stakes deployment

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

domains, dangerous emergent behaviors or vulnerabilities, or harmful misuses and abuses (Map 5.1).

- **Use red teams and adversarial testing** as part of extensive interaction with GPAI/foundation models to identify dangerous capabilities, vulnerabilities**,** or other emergent properties of such systems (Measure 1.1).
- **Track important identified risks** (e.g., vulnerabilities from data poisoning and other attacks or objectives mis-specification), even if they cannot yet be measured (Measure 1.1 and Measure 3.2).
- **Implement risk-reduction controls as appropriate** throughout a GPAI/foundation model's lifecycle, such as independent auditing, incremental scale-up, red teaming, and other steps (Manage 1.3, Manage 2.3, and Manage 2.4).
- **Incorporate identified AI system risk factors, and circumstances that could result in impacts or harms, into reporting and engagement with internal and external stakeholders** (e.g., to downstream developers, regulators, users, impacted communities, etc.) on the AI system as appropriate, e.g., using model cards, system cards, and other transparency mechanisms (Govern 4.2).

# 3. Topic Table

Table 1 maps risk management topics to the AI RMF core functions and subcategories. Subcategories that are referenced as "high-priority risk management steps" have been highlighted for ease of reference. Use this table to identify the subcategories relevant to specific core functions and topics.

Topics:

- **Context and Impact:** Practices and considerations that support identifying, understanding, and managing context-based impacts of GPAI/foundation model risks.
- **Transparency and Documentation:** Practices and considerations that support making a GPAI/foundation model's functionality, decision-making processes, and limitations clear and understandable to stakeholders.
- **Community Engagement:** Practices and considerations that support active collaboration with diverse stakeholders, including researchers, GPAI/foundation model developers and deployers, policymakers, affected communities, and the public.
- **Organizational Policy:** Practices and considerations that support establishing policies and guidelines for validity, reliability, safety, security, resilience, fairness, explainability, responsibility, and accountability throughout the AI lifecycle.
- **Testing and Evaluation:** Practices and considerations that support the testing and evaluation of GPAI/foundation model performance, capability, reliability, safety, and security. This also includes the validation of the methods.

**TABLE 1: TOPIC TABLE**

| | Govern | Map | Measure | Manage |
|---|---|---|---|---|
| **Context and Impact** | Govern 1.1<br>Govern 1.5 | Map 1.1<br>Map 1.3<br>Map 5.1 | Measure 2.11<br>Measure 2.12 | |
| **Transparency and Documentation** | Govern 2.1<br>Govern 4.2 | Map 2.2 | Measure 2.8<br>Measure 2.9<br>Measure 2.10 | Manage 1.3 |
| **Community Engagement** | Govern 3.1<br>Govern 4.2<br>Govern 5.1 | Map 5.2 | Measure 4.1 | Manage 1.3 |
| **Organizational Policy** | Govern 1.5<br>Govern 2.1 | Map 1.5<br>Map 5.1 | Measure 1.3<br>Measure 2.7<br>Measure 3.2 | Manage 1.1<br>Manage 1.3<br>Manage 2.4 |
| **Testing and Evaluation** | | | Measure 1.1<br>Measure 2.7<br>Measure 3.1 | |

# 4. Guidance

The high-level guidance provided below includes the "high-priority risk management steps" described in the previous section, as well as other unique perspectives and recommendations from across the AI RMF subcategories that may additionally be prioritized by organizations. Guidance that is one of the "high-priority risk management steps" is highlighted for ease of reference.

## 4.1 GOVERN

*Table 2 outlines priority guidance provided in the Profile under the "Govern" core function.*

*Note: These have been condensed for inclusion in the quick guide. Please refer to the full Profile for greater detail and context, including sources and resources. Guidance that is one of the "high-priority risk management steps" is highlighted for ease of reference.*

**TABLE 2: QUICK GUIDANCE FOR NIST AI RMF GOVERN**

| *Guidance* | *Subcategory* |
|---|---|
| **Assess the extent to which activities would fall under GPAI/foundation model-related laws or regulations**. (See, e.g., AI policy trackers.) | Govern 1.1 |
| **Identify GPAI/foundation model impacts** (including to human rights) and risks (including potential uses, misuses, and abuses), starting from an early AI lifecycle stage and repeating frequently through new lifecycle phases or as new information becomes available. | Govern 1.5 |
| **Revisit use and misuse case identification** at key intended milestones, or at periodic intervals (e.g., at least annually), whichever comes first. | Govern 1.5 |
| GPAI/foundation model developers should **take responsibility for risk assessment and risk management** tasks for which they have, or reasonably believe they might have, access to information, capability, or opportunity to develop capability sufficient for constructive action, or that is substantially greater than others in the value chain. | Govern 2.1 |
| **Make as much information transparent and available** on AI risk factors, incidents (including near-miss incidents), knowledge limits, etc., as reasonably possible to all audiences. | Govern 2.1 |
| As part of staffing to identify potential high-impact scenarios for GPAI/foundation models, **broaden the team as appropriate** to include, for example, social scientists and historians who can provide additional perspective on structural or systemic risks that could emerge from interactions between an AI system and other societal-level systems. | Govern 3.1 |
| **Incorporate identified AI system risk factors**, and circumstances that could result in impacts or harms, into **engagement** with internal and external stakeholders. | Govern 4.2 |

| *Guidance* | *Subcategory* |
|---|---|
| **Report identified AI system risk factors**, and circumstances that could result in impacts or harms to individuals (including impacts to health, safety, well-being, or fundamental rights), to organizations, to groups or communities, and to society (including risks to critical infrastructure, economic or national security, democratic institutions, or the environment). Also report additional identified factors that could lead to severe or catastrophic consequences for society, such as the potential for correlated robustness failures or other systemic risks across high-stakes application domains, potential for correlated bias across a large fraction of a society's population, potential for many high-impact uses or misuses beyond an originally intended use case, or other factors. | Govern 4.2 |
| GPAI/foundation model developers and deployers should **integrate independent feedback** from those external to the team that develops or deploys a model. Types of external feedback that should be utilized where appropriate include: deliberation with impacted communities, including people involved with the human labor and training of GPAI/foundation models (such as data annotators and content reviewers), people whose work is "scraped" for training purposes (such as artists and authors), intended users, and people whose livelihoods are altered by the use of the system. Feedback should also be gathered through independent auditing throughout the AI lifecycle; bug bounty and bias bounty programs; red-teaming; and feedback channels with users or impacted individuals or communities, including appeal and redress mechanisms. | Govern 5.1 |

## 4.2 MAP

*Table 3 details priority guidance provided in the Profile under the "Map" core function.*

*Note: These have been condensed for inclusion in the Quick Guide. Please refer to the full Profile for greater detail and context including sources and resources. Guidance that is one of the "high-priority risk management steps" is highlighted for ease of reference.*

**TABLE 3: QUICK GUIDANCE FOR NIST AI RMF MAP**

| *Guidance* | *Subcategory* |
|---|---|
| **Identify reasonably foreseeable uses, misuses, or abuses** for a GPAI/foundation model, beyond any originally intended use cases (or in the absence of a specific intended purpose). These may include:<br>• Automated generation of disinformation, or of phishing-attack material;<br>• Aiding with proliferation of chemical, biological, or radiological weapons, or other weapons of mass destruction. This can include aiding in lab experiment design and troubleshooting, providing instructions on chemical and biological material acquisition, and the capability to "upskill" threat actors by advising on attack techniques;<br>• Discovery and exploitation of software vulnerabilities, cyber attack plan critiquing and assistance, and malware and virus creation, including viruses that evolve over time to evade detection; and<br>• Creation of violent, illegal, discriminatory, or harmful content, including non-consensual intimate imagery (NCII) or child sexual abuse material. | Map 1.1 |
| **Identify the goals and limitations of the data collection and curation** processes, and the implications for the resulting AI systems. Consider running data audits as part of the data management process. | Map 1.1 |

| *Guidance* | *Subcategory* |
|---|---|
| **Identify reasonably foreseeable potential impacts** of GPAI/foundation models, which can include, but are not limited to:<br>• Impacts to organizational operations, including potential loss of understanding or control over particular functions, or loss of trust;<br>• Impacts to organizational assets, including legal compliance costs arising from problems created for individuals;<br>• Impacts to individuals, including impacts to health, safety, well-being, or fundamental rights;<br>• Impacts to groups, including populations vulnerable to disproportionate adverse impacts or harms;<br>• Impacts to society, including damage to or incapacitation of a critical infrastructure sector; economic and national security; concentration and control of the power and benefits from AI technologies; dramatic shifts to the labor market and economic opportunities, including technological job displacement; threats to democratic institutions and quality of life; and polarization and extremism; and<br>• Impacts to the environment, including carbon emissions and use of natural resources. | Map 1.1 |
| **Identify potential and actual human rights impacts**. Ensure alignment with the Universal Declaration of Human Rights (UDHR), including:<br>• The right to non-discrimination and equality before the law (Article 2);<br>• The right to life and personal security (Article 3);<br>• The right to privacy and protection against unlawful governmental surveillance (Article 12);<br>• The rights to freedom of thought, conscience, and religious belief and practice; freedom of expression; and freedom to hold opinions without interference (Articles 18 and 19);<br>• The rights to freedom of association and the right to peaceful assembly (Articles 20 and 21); and<br>• The rights to decent work and to an adequate standard of living (Article 23 and 25). | Map 1.1 |
| **Consider the following questions** for an AI system: "What **objective** has been specified for the system, and what kinds of perverse behavior could be incentivized by optimizing for that objective?" Examples of AI systems with mis-specified objectives can include machine-learning algorithms for social-media content recommendation that learn to optimize user-engagement metrics by serving users with extremist content or disinformation. | Map 1.3 |
| **Set policies on unacceptable-risk thresholds** for GPAI/foundation model development and deployment to include prevention of risks with substantial probability of significant, severe, or catastrophic outcomes if inadequately mitigated. Unacceptable-risk thresholds can be based on quantitative metrics, qualitative characteristics, or a combination of both. They should be informed not only by the risk tolerance of the organization in question, but also by broadly recognized notions of unacceptable risks to users and impacted communities, society, and the planet (see, for example, guidance on The G7 Hiroshima Process International Code of Conduct for Organizations Developing Advanced AI Systems.) For GPAI/foundation models with potential for unknown emergent properties, especially frontier models, consider including a "margin of safety" or buffer between the worst plausible system failures and the unacceptable-risk thresholds. Similar approaches are common for safety engineering in other fields. | Map 1.5 |
| **Fully scoping and understanding knowledge limits** of increasingly general-purpose AI systems is difficult. However, clear documentation and communication of their knowledge limits is very important, given the large number of potential uses of these AI systems. GPAI/foundation model developers should describe or list (and provide examples of) uses that would exceed a system's knowledge limits, as well as uses that would be appropriate given the system's knowledge limits. This information should be clearly featured in system documentation for downstream developers, users, and others as appropriate. | Map 2.2 |

| *Guidance* | *Subcategory* |
|---|---|
| **Prioritization of GPAI/foundation model risks** and potential impacts should include consideration of the magnitude of potential impacts, not just their likelihood. This is particularly important for any potential impacts with irreversible effects and catastrophic magnitude. Potential for such impacts can be more likely for GPAI/foundation models than for many other types of AI, because GPAI/foundation models are often more likely to have relatively greater capabilities, scale of deployment, and other factors leading to high impact. | Map 5.1 |
| **Factors that could lead to significant, severe, or catastrophic harms** to individuals, groups, organizations, and society can include:<br>• Correlated bias across large numbers of people or a large fraction of a group or society's population (e.g., resulting in systemic discrimination, exclusion, or violence);<br>• Impacts to societal trust or democratic processes, for example through large-scale manipulation of the information ecosystem;<br>• Correlated robustness failures across multiple high-stakes application domains such as critical infrastructure;<br>• Potential for high-impact misuses and abuses beyond originally intended use cases, including to assist in code generation for cybersecurity threats or to create or be used as destructive weapons, such as cyberweapons, lethal autonomous weapons, bio-weapons, or other significant military applications;<br>• Potential for large harms from mis-specified objectives or mis-generalized goals (e.g., using over-simplified or short-term metrics as proxies for desired longer-term outcomes);<br>• Ability to directly cause physical harms, e.g., via robotics motor control;<br>• Potential for socioeconomic risk and labor market disruption, including job automation and worsening inequality;<br>• Capability for AI models to manipulate or deceive humans into taking harmful actions in the world;<br>• AI systems that could recursively improve their capabilities by modifying their algorithms or architectures through code generation (e.g., from OpenAI Codex or DeepMind AlphaCode), neural architecture search, etc.;<br>• Adaptive models that might be difficult to control in real time, e.g., in response to the coordinated manipulation attacks, such as the attacks on the Microsoft Tay chatbot in 2016;<br>• Agentic systems, i.e. systems that in effect choose or take actions in a goal-directed fashion;<br>• Ability to employ outbound communication and influence channels, such as to post information to the Web via HTTP POST requests or functionally equivalent means (e.g., some types of plugins);<br>• Ability to escape a sandbox and replicate on another computational system, either via hacking, social engineering, or using other exploits;<br>• Sandbagging, i.e., strategically underperforming on model evaluations, including but not limited to password-locking or password-unlocking key capabilities; and<br>• Situational awareness, including a system being able to recognize that it is an AI, having knowledge about its capabilities and limitations, and knowing whether it is running in a test or deployment environment, which could, for example, enable a model to learn about the idea of jailbreaks from pre-training and utilize it when being evaluated for safety by a reward model. | Map 5.1 |
| GPAI/foundation model developers should **implement mechanisms to support regular engagement with relevant AI actors**, given the high likelihood and high potential impact of unanticipated negative impacts. These can include support for incident reporting, complaint and redress mechanisms, independent auditing, and protection for whistleblowers. | Map 5.2 |

## 4.3 MEASURE

*Table 4 outlines priority guidance provided in the Profile under the "Govern" core function.*

*Note: These have been condensed for inclusion in the Quick Guide. Please refer to the full Profile for greater detail and context including sources and resources. The "high-priority risk management steps" have been highlighted for ease of reference.*

**TABLE 4: QUICK GUIDANCE FOR NIST AI RMF MEASURE**

| *Guidance* | *Subcategory* |
|---|---|
| **Do not ignore identified risks just because measurement would be difficult**, especially if the impacts could be severe or catastrophic. Measurements of identified risks are often more difficult for GPAI/foundation models than for smaller-scale or fixed-purpose AI systems, because of factors such as complexities, uncertainties, and emergent properties of GPAI/foundation models. For many factors it can be more appropriate to use qualitative assessment procedures. Plan to track and revisit identified risks, even if they cannot be measured quantitatively at this time. | Measure 1.1 |
| **Use red teams and adversarial testing** as part of extensive interaction with GPAI/foundation models to identify dangerous capabilities, vulnerabilities, or other emergent properties. Security vulnerabilities are typically inherent to currently available GPAI/foundation models, including vulnerabilities to prompt injection attacks. Red-teaming can identify these weaknesses, though they are currently difficult to protect against. For frontier models, characteristics that red teams should evaluate include: unacceptable-risk factors as outlined in guidance under Map 1.5 and high-impact harm factors as outlined in guidance under Map 5.1. Partner with one or more independent red-teaming organizations as appropriate, and consider the following:<br>• Protect proprietary or unreleased foundation model weights as appropriate during red-teaming to prevent unauthorized access or leaks of model weights.<br>• Grant red teams considerable independence and control over the scrutiny process.<br>• Grant red teams appropriate access to the final versions of foundation models before deployment.<br>• For foundation models that are planned for release with downloadable, fully open, or open-source access, as part of pre-release red-teaming, allow red teamers to appropriately test the extent to which RLHF or other mitigations would not be resilient to additional fine-tuning or other processes used by actors with direct access to a model's weights after open release.<br>• As part of criteria for use of benchmarks or other metrics for risk assessment purposes, and as part of communication of benchmarking results, clarify whether a specific benchmark directly measures a particular risk, such as security vulnerability to prompt injection; whether it indicates a capability that could be misused or abused, such as software code generation; or whether it measures another important aspect of risk. | Measure 1.1 |
| **Encourage independent researchers to test models** and share their findings by offering bug and bias bounties, and by providing **safe harbo**r for AI evaluation and red-teaming. | Measure 1.3 |

| *Guidance* | *Subcategory* |
|---|---|
| **Use information security measures to assess and assure model weight security** (specifically, integrity and confidentiality) as part of preventing misuse or abuse of models. This is particularly valuable for frontier models, for which public release of model weights could enable misuse with particularly high-consequence impacts. Foundation model developers should implement the NIST Cybersecurity Framework (NIST 2024a), or an approximate equivalent such as NIST SP 800-171 or ISO/IEC 27001, with at least the following security controls or approximate equivalents:<br>• For frontier models: High-value asset guidance (e.g., per NIST SP 800-171 and NIST SP 800-172), or high-impact system baseline per NIST SP 800-53B as an informative reference for the NIST Cybersecurity Framework, or approximate equivalent.<br>• For other foundation models: Moderate-impact system baseline guidance (e.g., per NIST SP 800-171), or moderate-impact system baseline per NIST SP 800-53B as an informative reference for the NIST Cybersecurity Framework, or approximate equivalent. | Measure 2.7 |
| **Consider the following as part of security evaluations of GPAI/foundation models**:<br>• Perform red-teaming and adversarial testing of security aspects of GPAI/foundation models.<br>• Check for backdoors, AI trojans, prompt injection vulnerabilities, etc. during testing/evaluation, especially for models trained on untrusted data from public sources with susceptibility to data poisoning.<br>• Engage in continuous monitoring, vulnerability disclosure, and bug bounty programs for GPAI/foundation models to identify novel security vulnerabilities.<br>• Track uncovered security vulnerabilities in other GPAI/foundation models, including open-source foundation models, which may be transferable to other models. | Measure 2.7 |
| **Document organizational transparency and disclosure mechanisms** to inform or allow users to check whether they are interacting with, or observing content created by, a generative AI system. | Measure 2.8 |
| It is critical to **ensure that users know how to interpret system behavior and outputs**, including the limitations of both the system and any explanations provided. However, **explainability and interpretability** are often extremely limited for LLMs and other GPAI/foundation models with deep-learning architectures. These systems can be inappropriate for applications requiring better explainability and interpretability. | Measure 2.9 |
| **Privacy** challenges for GPAI/foundation models include the issue that, after pre-training on large quantities of uncurated Web-scraped data or other sources containing personally sensitive data, some of that sensitive material in the training data can be revealed by user prompts. **Enable people to consent to the uses of their data and opt out of the uses of their data. Also notify users and impacted communities about privacy or security breaches.** | Measure 2.10 |
| **Evaluating for fairness and bias** in GPAI/foundation models should take into account the complexity of the numerous sources and types of bias that influence GPAI/foundation models (including from massive datasets as well as decisions about modeling, optimization, hardware, and testing) and should not, for example, focus only on narrow definitions of protected classes, which may overlook complexities of identity. | Measure 2.11 |
| **Perform environmental-impact assessments**. GPAI/foundation model developers should include estimating the environmental impact of large-scale ML model training. Assessment of environmental impacts is particularly important for GPAI/foundation models, for which model training typically has much larger environmental impacts than smaller-scale ML models. | Measure 2.12 |

| *Guidance* | *Subcategory* |
|---|---|
| **Identify or assess longer-term impacts**, or use longer time horizons (longer than would be typical for smaller-scale fixed-purpose AI systems), to reduce potential for surprise. Consider whether any risk assessment or impact assessment answers would change if assessing **longer-term time periods (e.g., beyond the next year)**. What additional impacts would you expect? Which impacts would you expect to have greater magnitude? Identify unintended potential future events that should trigger reassessment or other responses, and build them into risk registers and/or planning and implementation of relevant lifecycle stages. | Measure 3.1 |
| **Use appropriate mechanisms for tracking identified risks**, even if only characterizing them qualitatively and even if the risks are difficult to assess. Consider tracking identified risks (including difficult-to-assess risks) using a risk register. When developing frontier models with unprecedented capabilities, failure modes, and other emergent properties, it is especially valuable to use red teams and adversarial testing prior to deployment. Risk tracking should include ongoing monitoring of newly identified capabilities and limitations of deployed GPAI/foundation models. These efforts can include monitoring use of the models through APIs, and monitoring publications or online forums that discuss new uses of the models. | Measure 3.2 |
| For GPAI/foundation model developers, model "**users**" include downstream developers as well as the end users of applications built on GPAI/foundation models. Downstream developers typically have the most direct interactions with end users in particular deployment contexts. However, it can be valuable for upstream GPAI/foundation model developers to **provide mechanisms for feedback from end users or other AI actors**, as well as from downstream developers. | Measure 4.1 |

## 4.4 MANAGE

*Table 5 outlines priority guidance provided in the Profile under the "Govern" core function.*

*Note: These have been condensed for inclusion in the Quick Guide. Please refer to the full Profile for greater detail and context including sources and resources. The "high-priority risk management steps" have been highlighted for ease of reference.*

**TABLE 5: QUICK GUIDANCE FOR NIST AI RMF MANAGE**

| *Guidance* | *Subcategory* |
|---|---|
| **Consider the following when making go/no-go decisions**, especially on whether to proceed on major stages or investments for development or deployment of cutting-edge large-scale GPAI/foundation models:<br>• See guidance in this document under Map 1.3 on AI development objectives, especially: Consider potential for mis-specified AI system objectives, and consider what kinds of perverse behavior could be incentivized by optimizing for those objectives.<br>• See guidance in this document under Map 1.5 on organizational risk tolerances, especially: Set policies on unacceptable-risk thresholds for GPAI/foundation model development and deployment to include prevention of risks with substantial probability of inadequately mitigated catastrophic outcomes. | Manage 1.1 |
| For each identified potential use or misuse (or category of use or misuse) of an AI system, define and **communicate to key stakeholders whether any potential use cases (or categories of use cases) would be disallowed/unacceptable**, or would be treated as "high risk." | Manage 1.3 |

| *Guidance* | *Subcategory* |
|---|---|
| If model training requires obtaining data sets, **consider using only trusted training data** instead of uncurated scrapes from the Web. This can be valuable for multiple objectives, including reducing vulnerability to backdoor and data-poisoning attacks, and reducing unwanted bias and language toxicity. | Manage 1.3 |
| **Increase the amount of compute (computing power) spent training frontier model**s **only incrementally** (e.g., by not more than three times between each increment) as part of identification and management of risks of emergent properties. **Test frontier models** after each incremental increase of compute, data, or model size for model training. Probe, monitor, and stress cutting-edge GPAI/foundation models using detailed analysis processes (including or extending standard cybersecurity red-team methods) to achieve testing objectives, including testing for unintended toxic and harmful content and/or dangerous errors (e.g., inaccurate medical information), and identifying emergent properties such as new capabilities and failure modes. | Manage 1.3 |
| **Consider approaches to design, testing, and deployment that ensure that AI systems possess only the minimum necessary capabilities** for high-reliability operation — and not more capabilities. Consider methods of implementing the cybersecurity principle of least privilege. For example, consider using or extending typical "deny by default" or whitelisting methods to limit an AI system's privileges to the minimum necessary for access to information, communication channels, and action space. | Manage 1.3 |
| **Determine a strategy to safely and appropriately release the AI system**, with consideration of what protections might be necessary to prevent harm or misuse. | Manage 1.3 |
| When planning for GPAI/foundation model deployment, **plan on deployment with gradual, phased releases, and/or structured access** through an API or other mechanisms, with efforts to detect and respond to misuse or problematic anomalies. Such systems and infrastructure can also be useful for enforcing usage guidelines and minimizing risks of misuse. | Manage 2.4 |
| GPAI/foundation model developers that plan to release a model with open-weights or open-source access, where that model would be above, at, or near a foundation model frontier, should first **use a staged-release approach** (e.g., they should not release model parameter weights until after an initial closed-source or structured-access release where no substantial risks or harms have emerged over a sufficient time period with red teaming and other evaluations as appropriate). They also should not proceed to a final step of releasing model parameter weights until a sufficient level of confidence in risk management has been established, including for safety and societal risks and risks of misuse and abuse. Models above a foundation model frontier should be given the greatest amount of duration and depth of pre-release evaluations, as they are most likely to have dangerous capabilities or vulnerabilities, or other properties that can take some time to discover. Foundation model developers that release a foundation model's parameter weights, or that suffer a leak of model weights, will in effect be unable to decommission AI systems that others build using those released or leaked foundation model weights. | Manage 2.4 |


CLTC
Center for Long-Term
Cybersecurity
UC Berkeley