UC BERKELEY

CENTER FOR LONG-TERM CYBERSECURITY

AI Risk-Management Standards Profile for General-Purpose AI Systems (GPAIS) and Foundation Models

Version 1.0, November 2023

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

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

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.

AI Risk-Management Standards Profile for General-Purpose AI Systems (GPAIS) and Foundation Models

ANTHONY M. BARRETT,[†*] **JESSICA NEWMAN,**[†] **BRANDIE NONNECKE,**[††] **DAN HENDRYCKS,**[†††]
EVAN R. MURPHY,[†] **KRYSTAL JACKSON**[†]

† 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 (affiliation during main contributions to this work)
* Corresponding author: anthony.barrett@berkeley.edu

CLTC Center for Long-Term Cybersecurity UC Berkeley

AI SECURITY INITIATIVE

CITRIS AND THE BANATAO INSTITUTE | CITRIS POLICY LAB

ABSTRACT

Increasingly multi-purpose AI systems, such as cutting-edge large language models or other "general-purpose AI" systems (GPAI or GPAIS), "foundation models," generative AI, and "frontier models" (typically all referred to hereafter with the umbrella term GPAIS except where greater specificity is needed), can provide many beneficial capabilities but also risks of adverse events with profound consequences. This document provides risk-management practices or controls for identifying, analyzing, and mitigating risks of GPAIS. We intend this document primarily for developers of large-scale, state-of-the-art GPAIS; others that can benefit from this guidance include downstream developers of end-use applications that build on a GPAIS platform. This document facilitates conformity with or use of leading AI risk management-related standards, adapting and building on the generic voluntary guidance in the NIST AI Risk Management Framework and ISO/IEC 23894, with a focus on the unique issues faced by developers of GPAIS.

Contents

EXECUTIVE SUMMARY 1

1. INTRODUCTION AND OBJECTIVES 5

1.1 Key Terms 5

1.2 Background and Purpose of the Profile 6

1.3 Intended Audience and Users of the Profile 8

1.4 Benefits of the Profile 9

1.4.1 Benefits of the Profile to Developers of GPAIS and Foundation Models 9

1.4.2 Benefits of the Profile to Deployers, Evaluators, and Users 10

1.4.3 Benefits for Individuals, Society, and the Regulatory Community 10

1.5 Limitations and Challenges 11

2. OVERVIEW OF PROFILE COMPONENTS AND HOW TO USE PROFILE 13

2.1 Basics 13

2.2 Impact Areas, Harm Factors, and Trustworthiness Characteristics 14

2.3 High-Priority Risk Management Steps and Profile Guidance Sections 15

3. GUIDANCE 17

3.1 Guidance for NIST AI RMF Govern Subcategories 17

GOVERN 1: Policies, processes, procedures, and practices across the organization related to the mapping, measuring, and managing of AI risks are in place, transparent, and implemented effectively. 17

GOVERN 2: Accountability structures are in place so that the appropriate teams and individuals are empowered, responsible, and trained for mapping, measuring, and managing AI risks. 21

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.* 21

GOVERN 3: Workforce diversity, equity, inclusion, and accessibility processes are prioritized in the mapping, measuring, and managing of AI risks throughout the lifecycle. 23

GOVERN 4: Organizational teams are committed to a culture that considers and communicates AI risk. 24

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.* 25

GOVERN 5: Processes are in place for robust engagement with relevant AI actors. 27

GOVERN 6: Policies and procedures are in place to address AI risks and benefits arising from third-party software and data and other supply chain issues. 28

3.2 Guidance for NIST AI RMF Map Subcategories 29

MAP 1: Context is established and understood. 29

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.* 29

MAP 1.5: *Organizational risk tolerances are determined and documented.* 32

MAP 2: Categorization of the AI system is performed. 33

MAP 3: AI capabilities, targeted usage, goals, and expected benefits and costs compared with appropriate benchmarks are understood. 34

MAP 4: Risks and benefits are mapped for all components of the AI system including third-party software and data. 35

MAP 5: Impacts to individuals, groups, communities, organizations, and society are characterized. 37

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.* 37

3.3 Guidance for NIST AI RMF Measure Subcategories 40

MEASURE 1: Appropriate methods and metrics are identified and applied. 40

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.* 40

MEASURE 2: AI systems are evaluated for trustworthy characteristics. 44

MEASURE 3: Mechanisms for tracking identified AI risks over time are in place. 51

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.* 52

MEASURE 4: Feedback about efficacy of measurement is gathered and assessed. 53

3.4 Guidance for NIST AI RMF Manage Subcategories 54

MANAGE 1: AI risks based on assessments and other analytical output from the Map and Measure functions are prioritized, responded to, and managed. 54

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.* 54

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.* 55

MANAGE 2: Strategies to maximize AI benefits and minimize negative impacts are planned, prepared, implemented, documented, and informed by input from relevant AI actors. 58

MANAGE 2.3: *Procedures are followed to respond to and recover from a previously unknown risk when it is identified.* 58

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.* 59

MANAGE 3: AI risks and benefits from third-party entities are managed. 60

MANAGE 4: Risk treatments, including response and recovery, and communication plans for the identified and measured AI risks are documented and monitored regularly. 61

4. MAPPING OF PROFILE GUIDANCE TO KEY STANDARDS AND REGULATIONS 63

4.1 Mapping to ISO/IEC 23894 63

4.2 Preliminary Mapping to ISO/IEC FDIS 42001 65

4.3 Mapping to White House AI Commitments 65

GLOSSARY 66

Acronyms

Terms

APPENDICES 68

Appendix 1: Overview of Development Approach 68
Appendix 2: Key Criteria for Guidance 68
Appendix 3: Roadmap of Issues to Address in Future Versions of the Profile 69
Appendix 4: Retrospective test use of Profile draft guidance 71
Appendix 4A: Profile draft-guidance testing methodology and main results 71
Appendix 4B: Feasibility issues identified in First Full Draft of Profile 75
Appendix 4C: GPT-4 78
Appendix 4D: Claude 2 82
Appendix 4E: PaLM 2 85
Appendix 4F: Llama 2 88

ACKNOWLEDGMENTS 93

REFERENCES 94

Executive Summary

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

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

We intend this Profile document primarily for use by **developers of large-scale, state-of-the-art GPAIS**. For GPAIS developers, this Profile document facilitates conformity with or use of leading AI risk management-related standards, and aims to facilitate compliance with relevant regulations such as the forthcoming EU AI Act, especially for aspects related to GPAIS. (However, this Profile does not provide all guidance that may be needed for GPAIS applications in particular industry sectors or applications.) Others who can benefit from the use of this guidance include: downstream developers of end-use applications that build on a GPAIS platform; evaluators of GPAIS; and the regulatory community. This document can provide GPAIS deployers, evaluators, and regulators with information useful for evaluating the extent to which developers of such AI systems have followed relevant best practices. Widespread norms for using best practices such as in this Profile can help ensure developers of GPAIS can be competitive without compromising on practices for AI safety, security, accountability, and related issues. Ultimately, this Profile aims to help key actors in the value chains of increasingly general-purpose AI systems to achieve outcomes of maximizing benefits, and minimizing negative impacts, to individuals, communities, organizations, society, and the planet. That includes protection of human rights, minimization of negative environmental impacts, and prevention of adverse events with systemic or catastrophic consequences at societal scale.

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

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

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

- **Check or update, and incorporate, each of the following 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 GPAIS (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, and misuses or abuses for a GPAIS** (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 GPAIS could lead to significant, severe, or catastrophic impacts,** e.g., because of correlated failures or errors across high-stakes deployment domains, dangerous emergent behaviors or vulnerabilities, or harmful misuses and abuses (Map 5.1).

- **Use red teams and adversarial testing** as part of extensive interaction with GPAIS 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 GPAIS lifecycle, e.g., independent auditing, incremental scale-up, red-teaming, structured access or staged release, and other steps (Manage 1.3, Manage 2.3, and Manage 2.4).

- **Incorporate identified AI system risk factors, and circumstances that could result in impacts or harms, into reporting and engagement with internal and external stakeholders** (e.g., 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).

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

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

- Reasonably foreseeable impacts (Section 3.2, Map 1.1), including:
  - To individuals, including impacts to health, safety, well-being, or fundamental rights;
  - To groups, including populations vulnerable to disproportionate adverse impacts or harms; and
  - To society, including environmental impacts.

- Significant, severe, or catastrophic harm factors (Section 3.2, Map 5.1), including:
  - Correlated bias and discrimination;

  - » Impacts to societal trust or democratic processes;
  - » Correlated robustness failures;
  - » Capability to manipulate or deceive humans in harmful ways; and
  - » Loss of understanding and control of an AI system in a real-world context.

- AI trustworthiness characteristics (Section 3.4, Measure 2), including:
  - » Safety, reliability, and robustness (Measure 2.5, Measure 2.6);
  - » Security and resiliency (Measure 2.7);
  - » Accountability and transparency (Measure 2.8);
  - » Explainability and interpretability (Measure 2.9);
  - » Privacy (Measure 2.10); and
  - » Fairness and bias (Measure 2.11)

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

1. Introduction and Objectives

1.1 KEY TERMS

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

We use these key terms as follows. (For additional terms and acronyms, see the Glossary.)

- **General-purpose AI system (GPAI or GPAIS):** "An AI system that can accomplish or be adapted to accomplish a range of distinct tasks, including some for which it was not intentionally and specifically trained" (Gutierrez et al. 2022, p. 22).
  - » We treat **GPAIS as an umbrella term that also includes foundation models, frontier models, and generative AI**, except where we need to be more specific.
  - » Examples of GPAIS include unimodal generative AI systems (e.g., GPT-3) and multimodal generative systems (e.g., DALL-E), as well as reinforcement-learning systems such as MuZero and AI systems with emergent capabilities, but exclude fixed-purpose AI systems trained specifically for tasks such as image classification or voice recognition (Gutierrez et al. 2022).

- **Foundation model**: "Any model that is trained on broad data (generally using self-supervision at scale) that can be adapted (e.g., fine-tuned) to a wide range of downstream tasks" (Bommasani et al. 2021, p. 3).
  - » We treat **foundation models as a large-scale, high-capability subset of pretrained GPAIS**, trained on relatively large data sets, resulting in relatively large-size pretrained models with relatively broad or high levels of capabilities, often released in ways that result in large numbers of users.
  - » Examples of foundation models include GPT-4, Claude 2, PaLM 2, LLaMA 2, and others.

- **Frontier model**: A cutting-edge, state-of-the-art, or highly capable GPAIS or foundation model. (See, e.g., Ganguli, Hernandez et al. 2022, Anderljung, Barnhart et al. 2023, Microsoft 2023a.)
  - » We treat **frontier models as the largest-scale, highest-capability subset of GPAIS or foundation models,** typically with model size, training compute or data, or resulting

capabilities, above or near to industry-record thresholds. (See also "foundation model frontier" in the Glossary.)
  - » Examples of frontier models: As of July 2023, models at or near the industry frontier include GPT-4, Claude 2, and PaLM 2.[1]

- **Generative AI**: "Any AI system whose primary function is to generate content" (Toner 2023).
  - » We typically only use the term **"generative AI" to highlight issues specific to synthetic text (which can include software code), images, video, audio, or other synthetic media**. (In some other documents, "generative AI" is used in approximately the same way that we use the terms GPAIS or foundation model.)
  - » Examples of generative AI: "Typical examples of generative AI systems include image generators (such as Midjourney or Stable Diffusion), large language models (such as GPT-4, PaLM 2, or Claude 2), code generation tools (such as Copilot), or audio generation tools (such as VALL-E or resemble.ai)" (Toner 2023).

We intend our usage of the terms "general-purpose AI" or GPAIS, "foundation model," and "generative AI" to be broadly compatible with usage of the equivalent terms where applicable in the OECD classification framework (OECD 2022a, p. 64), draft EU AI Act, and October 30th, 2023 Biden Executive Order (White House 2023c), and the Hiroshima Process International Code of Conduct for Advanced AI Systems (G7 2023), though our focus in this document is primarily on the most broadly capable AI systems meeting the definitions for GPAIS and foundation models.[2]

1.2 BACKGROUND AND PURPOSE OF THE PROFILE

GPAIS and foundation models such as GPT-4, DALL-E 3, PaLM 2, Claude 2, and Llama 2 can serve as multi-purpose AI platforms underpinning many end-use applications. These increasingly powerful GPAIS are the focus of cutting-edge research. They also have several qualitatively distinct properties compared to the more common, narrower machine learning models, such

1 Several AI companies have committed to measures such as red teaming and public reporting of societal risks when developing and releasing models more powerful than GPT-4 or other models at the July 2023 industry frontier (White House 2023a).

2 The European Parliament amendments to the AI Act propose a definition for GPAIS as "an AI system that can be used in and adapted to a wide range of applications for which it was not intentionally and specifically designed" (EP 2023 Amendment 169 Article 3 paragraph 1 point 1d), a definition for the more capable and larger-scale subset of GPAIS identified as foundation models as "an AI system model that is trained on broad data at scale, is designed for generality of output, and can be adapted to a wide range of distinctive tasks" (EP 2023 Amendment 169 Article 3 paragraph 1 point 1c), and an implied definition of generative AI as a foundation model or other AI system "specifically intended to generate, with varying levels of autonomy, content such as complex text, images, audio, or video" (EP 2023 Amendment 399 Article 28b).

as potential to be applied to many sectors at once, potential large-scale societal, environmental, security, and economic impacts, and emergent properties that can provide unexpected beneficial capabilities but also unexpected risks of adverse events[3] (Bommasani et al. 2021, Weidinger et al. 2021, Wei et al. 2022). These properties complicate the ways in which general-purpose AI systems can be governed, though many AI experts encourage their inclusion in regulatory and risk management frameworks (see, e.g., Gebru et al. 2023). It can be appropriate to carry out more in-depth risk assessment with longer time horizons, at more points in the AI system life cycle, and to implement other, more extensive risk-mitigation controls, for GPAIS than for AI with more limited capabilities.

This document is designed to complement the broadly applicable guidance in the NIST AI Risk Management Framework, or AI RMF (NIST n.d.a), or a related AI risk management standard such as ISO/IEC 23894. This document provides an AI risk-management standards target Profile, with a set of risk-management practices or controls and target outcomes specifically for identifying, analyzing, and mitigating risks of GPAIS and foundation models. This cross-sectoral Profile addresses important underlying risks and early-development risks of such technologies in a way that does not rely on great certainty about each specific end-use application of the technology. We have developed this as a community Profile in a multi-stakeholder process with input and feedback on drafts from a range of stakeholders, including organizations developing large-scale GPAIS and foundation models, and other organizations across industry, civil society, academia, and government.

AI risk categories we aim to address with the guidance in this document include:

- Risks stemming from the large scale and reach of GPAIS, resulting from their frequent place in the AI value chain as foundation models that many other systems build on and rely upon.

- Risks of misuse and abuse of GPAIS, resulting from their lowering barriers for malicious activities such as generating disinformation.

- Risks of unexpected impacts of GPAIS, resulting from the emergent behaviors, vulnerabilities, and capabilities that are often found (and continue to be found) in state-of-the-art large-scale GPAIS.

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

Guidance in this Profile for GPAIS is based in part on examples of assessments and/or risk management controls already implemented by market leaders such as DeepMind, OpenAI, and Hugging Face. For example, OpenAI's 2019 announcement of GPT-2 included enumeration of several categories of potential misuse cases (OpenAI 2019a), which apparently informed OpenAI's decisions on disallowed/unacceptable use-case categories of applications based on GPT-3 (OpenAI 2020). DeepMind's 2021 announcement of their large language model Gopher and 2022 announcement of their multi-modal and multi-task "generalist agent" Gato also included consideration of potential misuse, safety risks, and mitigation (Rae et al. 2021; Weidinger et al. 2021; Reed et al. 2022). Hugging Face and BigScience's release of the BLOOM LLM included a Responsible AI License (RAIL) with usage restrictions disallowing various types of misuse (RAIL n.d., Contractor et al. 2022). The Partnership on AI has developed guidance on synthetic media, including on transparency and disclosure of generative AI outputs (PAI 2023a), and is also working to develop protocols for responsible deployment of foundation models (PAI 2023b, c). The newly created Frontier Model Forum has also announced plans to research and share best practices for development of highly capable foundation models or frontier models, including safety-related evaluations (Heath 2023). In addition, NIST has created a Generative AI Public Working Group, and plans to create a NIST AI RMF profile specifically on generative AI (NIST 2023d).[4]

Some of the material in this Profile is adapted directly from our related work in Section 4, or other sections, of Barrett et al. (2022). Some other material in Section 3 of this Profile consists of extended excerpts from the NIST AI RMF Playbook (NIST 2023b), highlighting the portions of the broadly applicable Playbook guidance that seem particularly valuable for GPAIS developers, in light of typical current GPAIS architectures and development practices.

1.3 INTENDED AUDIENCE AND USERS OF THE PROFILE

We intend this document primarily for use by **developers of large-scale, state-of-the-art general-purpose AI systems or GPAIS, foundation models, and generative AI systems**; others who can benefit from use of this guidance include **downstream developers of end-use applications that build on a GPAIS platform**.

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

4 Our Berkeley GPAIS and foundation model Profile effort is separate from, but aims to complement and inform the work of, other guidance development efforts such as the PAI Guidance for Safe Foundation Model Deployment and the NIST Generative AI Public Working Group.

- One of a few large-scale GPAIS platforms or foundation models. These AI systems (and especially the most broadly capable GPAIS) are the main focus of this Profile, with some corresponding costs for upstream GPAIS developers, but also corresponding risk-management benefits when employing the guidance in this Profile.

- A relatively narrow-purpose end-use application that builds on a GPAIS or foundation model. Some aspects of these end-use application AI systems are constructively addressed by parts of the guidance in this Profile. Costs to downstream developers building applications on GPAIS would likely be minimal when employing relevant guidance in this Profile.

- One of many small-scale or stand-alone narrow-purpose systems that do not fall under definitions for GPAIS, and are not within the scope of this Profile. We do not expect developers or deployers of these common AI systems to use this Profile for those AI systems, and thus we do not expect their costs to be substantially affected by this Profile.

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

- Open-source and open-access GPAIS development efforts, as well as closed-source GPAIS development; and

- Research projects, and other GPAIS that a GPAIS developer does not plan to make available to users outside the organization, as well as GPAIS that a GPAIS developer plans to put on the market.

1.4 BENEFITS OF THE PROFILE

1.4.1 Benefits of the Profile to Developers of GPAIS and Foundation Models

This Profile provides developers of GPAIS and foundation models with valuable risk-management best practices addressing their unique issues. For example, the Profile provides guidance on sharing of responsibilities between (a) upstream developers that create GPAIS and offer AI

platforms/APIs based on those AI systems in a manner that allows many different end uses, and (b) downstream developers that build upon the GPAIS platforms for specific end-use applications using upstream provider-supplied information that may not be customized for their own application area.

This document facilitates conformity with or use of leading AI risk management-related standards, adapting and building on the generic voluntary guidance in the NIST AI Risk Management Framework and ISO/IEC 23894, with a focus on the unique issues faced by developers of GPAIS. It also aims to facilitate compliance with relevant regulations, such as the forthcoming EU AI Act, especially for aspects related to GPAIS, foundation models, and generative AI. For example, this document could help fulfill expectations for transparency, risk management, audits, etc. as applicable to GPAIS and foundation models under the AI Act.[5]

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

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

This Profile can provide deployers, evaluators, and users of GPAIS with increased awareness of the risks of such AI systems and of best practices to use in addressing those risks. This document also can provide deployers, evaluators, and users of such AI systems with information useful for evaluating the extent to which developers of such AI systems have followed relevant best practices.

1.4.3 Benefits for Individuals, Society, and the Regulatory Community

Ultimately, this Profile aims to help key actors in the value chains of increasingly general-purpose AI systems to achieve outcomes of maximizing benefits, and minimizing negative impacts, to individuals, communities, organizations, society, and the planet. That includes protection of fundamental rights, minimization of negative environmental impacts, and prevention of adverse events with systemic or catastrophic consequences at societal scale. There are vital relationships between principles of fairness and protecting human rights, addressing risks to individuals and groups, and addressing large-scale systemic or catastrophic risks. Some types of risks to individuals or groups comprise significant, severe, or catastrophic risks via accumulation or cor-

5 At the time of writing this document, requirements for GPAIS under the EU AI Act are still in negotiations. Draft requirements for GPAIS under the Act reportedly include providing non-commercially sensitive information, and may also include risk and quality management requirements and external audits, among other obligations (Bertuzzi 2023a,b,c,d,e and EP 2023).

relation of risks across individuals. Managing risks of GPAIS should include appropriate protection of human rights, and consideration of populations vulnerable to disproportionate harms. Preventing catastrophe can also be an important part of preventing unfair outcomes; often the effects of catastrophe fall disproportionately on disadvantaged people. It is critical to ensure that communities that may use or be impacted by the AI systems are meaningfully involved throughout the AI lifecycle, with opportunities to provide feedback and report potential problems.

From a regulatory perspective, this document can be viewed as part of "soft law" norms and best practices that GPAIS developers and deployers would have incentives to follow as appropriate, and that regulators can consider when formulating relevant "hard law" regulations (see, e.g., Gutierrez et al. 2021).[6] We also aim to provide mapping to, and harmonization with, relevant standards (e.g., ISO/IEC 23894 and ISO/IEC 42001) and regulations (e.g., the EU AI Act). This would help to set norms for GPAIS risk-management practices and conformity across regulatory regimes.

1.5 LIMITATIONS AND CHALLENGES

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

Another limitation is the relatively nascent state of best practices for developers of GPAIS. We have based our guidance on available literature, demonstrated industry practices, stakeholder input and feedback, and ultimately our own judgment. However, we expect that best practices in this area will continue to evolve substantially. At minimum, we expect that such further evolution will provide more detailed resources in a number of areas, which we aim to incorporate in later versions of this guidance, e.g., in annual updates.

6 As a related example, the US National Telecommunications and Information Administration (NTIA) will be making AI accountability policy recommendations that could include US government procurement mandates for audits (Fried 2023). NTIA is including LLMs and other GPAIS in its considerations (NTIA 2023).

Challenges in this guidance include tradeoffs between risks and benefits, and even between different sets of risks. One of the most challenging areas is open-source development and release, or closely related release strategies such as downloadable and fully open-access release of model weights. There is great value in open-source software and various forms of transparency and access to AI systems, including for helping to ensure the safety and security of an AI system's intended users. However, providing direct access to a model's weights also can increase some types of risks, including risks of malicious misuse. GPAIS developers that publicly release the model parameter weights for their GPAIS with downloadable, fully open, or open-source access to their models, and other GPAIS developers that suffer a leak of model weights, will in effect be unable to shut down or decommission GPAIS that others build using those model weights. This is a consideration that should be weighed against the benefits of open-source models, especially for the largest-scale and most broadly capable models that pose the greatest risks of enabling severe harms, including from malicious misuse to harm the public. Many of the benefits of openness, such as review and evaluation from a broader set of stakeholders and greater ability to use a model, are possible to support either through transparency, engagement, or other openness mechanisms that do not require a model's parameter weights to become downloadable or open-sourced, or through smaller-scale and less broadly capable open-source models. Thus, our profile guidance includes many transparency and access provisions, including under Govern 4.2 on reporting to 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, or system cards, and other transparency mechanisms. Another important part of our profile guidance (under Manage 2.4) is that GPAIS and foundation model developers that plan to provide downloadable, fully open, or open-source access to their models should first use a staged-release approach (e.g., not releasing parameter weights until after an initial closed-source or structured-access release where no substantial risks or harms have emerged over a sufficient time period), and should not proceed to a final step of releasing model parameter weights until a sufficient level of confidence in risk management has been established, including for safety risks and risks of misuse and abuse. (That level of confidence in safety would be particularly difficult to appropriately establish for the largest-scale or most capable models, and they should be given the greatest duration and depth of pre-release evaluations, as they are the most likely to have dangerous capabilities or other emergent properties that can take some time to discover.) We believe this overall approach provides actionable guidance to address some of the greatest risks to the public associated with open-sourcing powerful AI models, while providing valuable transparency mechanisms, and without prohibiting responsible open-sourcing of AI models.

2. Overview of Profile Components and How to Use Profile

2.1 BASICS

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

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

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

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

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

Our usage of the terms "should" and "can" in the guidance in Section 3 of this document is as follows: "should" indicates our recommendation and "can" indicates something is possible.[7]

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

2.2 IMPACT AREAS, HARM FACTORS, AND TRUSTWORTHINESS CHARACTERISTICS

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

- Reasonably foreseeable impacts (Section 3.2, Map 1.1), including:
  - To individuals, including impacts to health, safety, well-being, or fundamental rights;
  - To groups, including populations vulnerable to disproportionate adverse impacts or harms; and
  - To society, including environmental impacts.

- Significant, severe, or catastrophic harm factors (Section 3.2, Map 5.1), including:
  - Correlated bias and discrimination;
  - Impacts to societal trust or democratic processes;
  - Correlated robustness failures;
  - Capability to manipulate or deceive humans in harmful ways; and
  - Loss of understanding and control of an AI system in a real world context (e.g., ability to escape a sandbox and replicate on another computational system).

- AI trustworthiness characteristics (Section 3.4, Measure 2), including:
  - Safety, reliability, and robustness (Measure 2.5, Measure 2.6);
  - Security and resiliency (Measure 2.7);
  - Accountability and transparency (Measure 2.8);
  - Explainability and interpretability (Measure 2.9);
  - Privacy (Measure 2.10); and
  - Fairness and bias (Measure 2.11).

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

2.3 HIGH-PRIORITY RISK MANAGEMENT STEPS AND PROFILE GUIDANCE SECTIONS

Users of this Profile should place high priority on the following risk management steps and corresponding Profile guidance sections.[8] (Appropriately applying the Profile guidance for the following steps should be regarded as the baseline or minimum expectations for users of this Profile; users of this Profile can exceed the minimum expectations by also applying guidance in other sections.)

- **Check or update, and incorporate, each of the following 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 GPAIS (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 GPAIS** (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).

8 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.

- **Identify whether a GPAIS could lead to significant, severe, or catastrophic impacts,** e.g., because of correlated failures or errors across high-stakes deployment domains, dangerous emergent behaviors or vulnerabilities, or harmful misuses and abuses (Map 5.1).

- **Use red teams and adversarial testing** as part of extensive interaction with GPAIS 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 GPAIS lifecycle, e.g., 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).

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

3. Guidance

Broadly speaking, all areas of current NIST AI RMF guidance (NIST 2023a, 2023b) seem at least partly applicable for GPAIS. However, for such AI systems, the activities and outcomes for some categories or subcategories of NIST AI RMF guidance seem higher priority than others. In the following, we have included a number of excerpts from the NIST AI RMF Playbook (NIST 2023b) that seem particularly valuable for GPAIS developers, given current typical GPAIS architectures and development practices. NIST AI RMF Playbook excerpts in the following are in *italics font*, and are preceded by statements of the form "In the NIST AI RMF Playbook guidance for __, particularly valuable action and documentation items for GPAIS include __."

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

3.1 GUIDANCE FOR NIST AI RMF GOVERN SUBCATEGORIES

Table 1: Guidance for NIST AI RMF Govern Subcategories

Govern Category or Subcategory	Applicability and supplemental guidance for GPAIS	Resources
Govern 1: Policies, processes, procedures, and practices across the organization related to the mapping, measuring, and managing of AI risks are in place, transparent, and implemented effectively.		
Govern 1.1: Legal and regulatory requirements involving AI are understood, managed, and documented.	The legal and regulatory environment for GPAIS and generative AI is evolving quickly and will require regular assessment for continued compliance. GPAIS developers, deployers, and users should assess the extent to which their activities would fall under GPAIS-related laws or regulations, such as: • The forthcoming EU AI Act, which reportedly is likely to include regulatory requirements for GPAIS, foundation models, and generative AI. At a minimum, GPAIS (or at least one or more subsets of GPAIS identified as foundation models) seem likely to be subject to requirements for transparency, and for assessing, mitigating, and documenting several types of reasonably foreseeable risks[9] (Bertuzzi 2023b,c,d,e). • Numerous bills introduced in the United States related to generative AI, at both the federal and state levels, including in California, Massachusetts, and New York.	NIST (2023b) On copyright and fair use: Henderson et al. (2023) Samuelson (2023) On the EU AI Act: Schuett (2023) Bommasani et al. (2023)

9 We aim to provide mapping of profile guidance to relevant clauses of the EU AI Act after its finalization, in Section 4 of a future version of this Profile.

Govern Category or Subcategory	Applicability and supplemental guidance for GPAIS	Resources
(continued)	• China's Generative AI Regulation, which went into effect in August 2023 and introduced requirements related to acceptable generated content, using legitimate and legal sources for data training, obtaining consent for personal information processing, and submitting certain generative AI services for a security assessment, among other measures. Issues related to copyright, data protection, and privacy rights are also particularly relevant to GPAIS that are trained on large swaths of the internet. Many professional artists and writers oppose the use of their work as training data for GPAIS, and numerous copyright lawsuits are now underway in the United States (see, e.g, Samuelson 2023).	
Govern 1.2: The characteristics of trustworthy AI are integrated into organizational policies, processes, procedures, and practices.	The characteristics of trustworthy AI, described in the NIST AI RMF, include: valid and reliable, safe, secure and resilient, accountable and transparent, explainable and interpretable, privacy-enhanced, and fair with harmful bias managed. For GPAIS, there are some unique or particularly important considerations related to ensuring the characteristics of trustworthy AI are integrated into organizational policies, processes, procedures, and practices (Newman 2023; Wang et al., 2023). Some of these are mentioned below; see also the more detailed considerations and guidance throughout this document: Valid and reliable: • E.g., improve predictability, review dependencies on external parties, assess quality of training data, and train operators of the system to exercise oversight and avoid overconfidence in the system. Safe: • E.g., establish reliable technical and procedural controls, re-evaluate safety regularly, assess shifts over time, and report incidents and adverse impacts. Fair with harmful bias managed: • E.g., engage with impacted communities, test for biased or discriminatory outputs, review impacts on human rights and wellbeing, assess accessibility of user interface, and assess how to equitably distribute benefits. Secure and Resilient: • E.g., assess robustness in novel environments, establish protections against adversarial attacks, and establish a coordinated policy to encourage responsible vulnerability research and disclosure. Explainable and Interpretable: • E.g., ensure users know how to interpret system behavior and outputs, including limitations. Privacy-enhanced: • E.g., enable people to consent to the uses of their data and opt out of the uses of their data, and notify users about privacy and security breaches. Accountable and Transparent: • E.g., determine a publication/release strategy, inform users when they are interacting with the AI system or viewing AI-generated content, allow people to opt out, support independent third-party auditing and evaluation, and provide redress to people who are negatively affected.	NIST (2023b) Newman (2023) Wang et al. (2023)

Govern Category or Subcategory	Applicability and supplemental guidance for GPAIS	Resources
Govern 1.3: Processes, procedures, and practices are in place to determine the needed level of risk management activities based on the organization's risk tolerance.	GPAIS often can have greater impacts or pose greater risks than smaller or less capable AI systems due to their potential use in many different downstream applications. Therefore, for GPAIS it would often be appropriate to make GPAIS risk assessment and management a higher priority, and devote more resources, as compared with lower-capability and lower-impact AI systems. (See also the material in this document under Map 1.1 and Map 5.1 for related guidance on GPAIS impact assessment, including on impact identification and impact magnitude rating, and under Map 1.5 on risk tolerance, including on setting unacceptable risk thresholds.)	NIST (2023b)

Govern 1.4: The risk management process and its outcomes are established through transparent policies, procedures, and other controls based on organizational risk priorities.	In the NIST AI RMF Playbook guidance for Govern 1.4, particularly valuable action and documentation items for GPAIS include:
	• *Establish and regularly review documentation policies that, among others, address information related to:*
	  ◦ *Expected and potential risks and impacts*
	  ◦ *Assumptions and limitations*
	  ◦ *Description and characterization of training data*
	  ◦ *Testing and validation results (including explanatory visualizations and information)*
	  ◦ *Down- and up-stream dependencies*
	  ◦ *Plans for deployment, monitoring, and change management*
	  ◦ *Stakeholder engagement plans*
	• *Establish policies and processes regarding public disclosure of the use of AI and risk management material such as impact assessments, audits, model documentation and validation and testing results.*
	• *Document and review the use and efficacy of different types of transparency tools and follow industry standards at the time a model is in use.*
	
	(When considering disclosure of risk management material such as impact assessments, audits, model documentation, and validation and testing results, see also the material under Govern 4.2 in this document for related guidance on documentation and communication.)
		NIST (2023b)
		Bender et al. (2022)
		Gebru et al. (2021)
		Mitchell et al. (2019)

Govern 1.5: Ongoing monitoring and periodic review of the risk management process and its outcomes are planned and organizational roles and responsibilities clearly defined, including determining the frequency of periodic review.	**Plan to identify GPAIS impacts (including to human rights) and risks (including potential uses, misuses, and abuses), starting from an early AI lifecycle stage and repeatedly through new lifecycle phases or as new information becomes available.** This is particularly important for GPAIS, which can have large numbers of uses, risks, and impacts, including from emergent capabilities and vulnerabilities.
	• **On GPAIS lifecycle and when to assess risks:**
	  ◦ For larger machine learning models, iterations are often slower than typical Agile sprints. For larger models, the pipeline is often to pretrain a model, analyze, customize, reanalyze, customize differently, etc., then deploy and monitor, then decommission. (Here we use "analyze" as a shorthand for probing, stress testing, red teaming, monitoring in simulated environments, etc.)
	  ◦ On red teaming, see e.g., Ganguli, Lovitt et al. (2022), and guidance in this document under Measure 1.1.
	• All the relevant parties, especially researchers involved in the R&D process, should have some minimal knowledge on the risks of GPAIS or be taught about such risks upon their inclusion on an advisory team.
		Barrett et al. (2022)
		PAI (2023a)
		NIST (2023b)

Govern Category or Subcategory	Applicability and supplemental guidance for GPAIS	Resources

(continued)	• For larger models or close-to-frontier models, "Map" activities to identify risks should also happen after model training to incorporate developer findings about a model's capabilities post-training, not just before model training.
	• **On identifying potential uses, misuses, and abuses of a GPAIS:**
	  ◦ Identify potential use cases during early stages of your AI system lifecycle, such as the plan and design stages, at minimum.
	  ◦ Identify misuse or abuse cases during all major stages of your AI system lifecycle (or approximate equivalents in Agile/iterative development sprints), such as: plan, data collection, design, train/build/buy, test and evaluation, deploy, operate and monitor, and decommission.
	  ◦ Revisit use and misuse case identification at key intended milestones, or at periodic intervals (e.g., at least annually), whichever comes first.
	  ◦ Create a plan for ongoing use case identification and categorization to extend identified uses, misuses, and abuses, based on information gained continuously from sources such as:
	  » Downstream user and developer exploration of the AI system.
	  » API misuse and abuse monitoring.
	• **When making go/no-go decisions**, especially on whether to proceed on major stages or investments for development or deployment of cutting-edge large-scale GPAIS, see guidance in this document under Manage 1.1.
	  ◦ It can be valuable to revisit risk assessment at these intervals, especially prior to beginning a new frontier-model training run. At or near a foundation model frontier, it would be particularly important to obtain and integrate new information on emergent properties of frontier models before incurring the expenditures and risks of the next big training run.
	
	In the NIST AI RMF Playbook guidance for Govern 1.5, particularly valuable action and documentation items for GPAIS include:
	• *Establish policies to allocate appropriate resources and capacity for assessing impacts of AI systems on individuals, communities and society.*
	• *Establish policies and procedures for monitoring and addressing AI system performance and trustworthiness, including bias and security problems, across the lifecycle of the system.*
	• *Establish policies for AI system incident response, or confirm that existing incident response policies apply to AI systems.*
	• *Establish policies to define organizational functions and personnel responsible for AI system monitoring and incident response activities.*
	• *Establish mechanisms to enable the sharing of feedback from impacted individuals or communities about negative impacts from AI systems.*
	• *Establish mechanisms to provide recourse for impacted individuals or communities to contest problematic AI system outcomes.*	

Govern 1.6: Mechanisms are in place to inventory AI systems and are resourced according to organizational risk priorities.	(No supplemental guidance, beyond the broadly applicable guidance in the NIST AI RMF Playbook.)	NIST (2023b)

Govern Category or Subcategory	Applicability and supplemental guidance for GPAIS	Resources

Govern 1.7: Processes and procedures are in place for decommissioning and phasing out AI systems safely and in a manner that does not increase risks or decrease the organization's trustworthiness.	Open-source and fully open-access GPAIS developers that publicly release the model parameter weights for their GPAIS, and other GPAIS developers that suffer a leak of model weights, will in effect be unable to decommission GPAIS that others build using those model weights. (See also guidance in this document under Manage 2.4, recommending structured access or staged release approaches, including for foundation model developers that plan to fully open-source their models.)
	
	In the NIST AI RMF Playbook guidance for Govern 1.7, particularly valuable action and documentation items for GPAIS include:
	• *Establish policies for decommissioning AI systems. Such policies typically address:*
	  ◦ *User and community concerns, and reputational risks.*
	  ◦ *Business continuity and financial risks.*
	  ◦ *Up and downstream system dependencies.*
	  ◦ *Regulatory requirements (e.g., data retention).*
	  ◦ *Potential future legal, regulatory, security or forensic investigations.*
	  ◦ *Migration to the replacement system, if appropriate.*
	• *If anyone believes that the AI no longer meets this ethical framework, who will be responsible for receiving the concern and as appropriate investigating and remediating the issue? Do they have authority to modify, limit, or stop the use of the AI?*
		NIST (2023b)

Govern 2: Accountability structures are in place so that the appropriate teams and individuals are empowered, responsible, and trained for mapping, measuring, and managing AI risks.		

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.	**Regarding roles and responsibilities across a GPAIS value chain:**
	• **GPAIS developers should be responsible 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**, such as:
	  ◦ Assessing and mitigating early-stage development risks, including for AI research projects and AI systems that the organization does not plan to make available to others.
	  ◦ Testing and documentation that require direct access to training data or the AI system, such as on knowledge limits and dangerous capabilities.
	  ◦ Identifying reasonably foreseeable uses, misuses, and abuses of the AI system.
	  ◦ Implementing appropriate precautions to prevent or mitigate identified potential misuses or abuses.[10]
	  ◦ Making necessary information available to downstream developers and deployers building on base-model or GPAIS platforms, and to independent auditors or others as appropriate (e.g., to enable third-party auditability).
	  » Make as much information available on AI risk factors, incidents (including near-miss incidents), knowledge limits, etc., as reasonably possible to all audiences.[11]
		Barrett et al. (2022)
		NIST (2023b)
		Schuett (2022)
		Schuett (2023)

10 See also Manage 1.3 guidance on defining and communicating to key stakeholders whether any potential use cases would be disallowed/unacceptable.

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

Govern Category or Subcategory	Applicability and supplemental guidance for GPAIS	Resources
(continued)	» Provide additional information to downstream and end-use application developers and deployers as appropriate to meet their risk management needs. • **Downstream developers and deployers** of end-use applications built on GPAIS **should be responsible 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**, such as: ◦ Establishing specific context for their intended end-use application(s), and applying risk management processes appropriate for that specific context. ◦ Utilizing information provided by the upstream provider of an AI system platform, and requesting additional information as needed. ◦ Reporting to the upstream provider, and considering reporting to others such as information sharing and analysis organizations (ISAOs) or regulators as appropriate, any critical GPAIS vulnerabilities, biases, incidents (including near-miss incidents), etc., that would have high impacts on other downstream developers or deployers. • Downstream developers and deployers extending GPAIS (e.g., via fine-tuning training on data curated by the downstream developer) should also consider applying guidance for upstream developers (e.g., on testing and documentation that require direct access to fine-tuning training data) for any substantial extensions of the underlying platform AI systems. Fine-tuned versions of the underlying platform systems often have capabilities that underlying platform systems do not. Regarding roles and responsibilities for accountability within a single GPAIS developer or deployer organization, consider implementing "Three Lines of Defense" or 3LoD (Schuett 2022): • Roles can include: 1. Research team as the first line, ultimately the Head of Research or equivalent; 2. Risk management team as the second line, ultimately chief risk officer (CRO) or equivalent; this can also include the legal and compliance team, technical safety team, and security team; and 3. Internal audit as third line, ultimately chief audit executive (CAE); this can also include the ethics board. • Reporting responsibilities can include: 1. First line reports to CEO; 2. Second line reports to CEO; and CRO reports to the board risk committee; and 3. Third line reports to the board of directors or the board audit committee; the CAE is often part of the board audit committee. In the NIST AI RMF Playbook guidance for Govern 2.1, particularly valuable action and documentation items for GPAIS include: • *Establish policies that define the AI risk management roles and responsibilities for positions directly and indirectly related to AI systems, including, but not limited to - Boards of directors or advisory committees - Senior management - AI audit functions - Product management - Project management - AI design - AI development - Human-AI interaction - AI testing and evaluation - AI acquisition and procurement - Impact assessment functions - Oversight functions.*	

Govern Category or Subcategory	Applicability and supplemental guidance for GPAIS	Resources
(continued)	• *Establish policies that promote regular communication among AI actors participating in AI risk management efforts.* • *Establish policies that separate management of AI system development functions from AI system testing functions, to enable independent course-correction of AI systems.*	
Govern 2.2: The organization's personnel and partners receive AI risk management training to enable them to perform their duties and responsibilities consistent with related policies, procedures, and agreements.	In the NIST AI RMF Playbook guidance for Govern 2.2, particularly valuable action and documentation items for GPAIS include: • *Ensure that trainings comprehensively address technical and socio-technical aspects of AI risk management.* • *Define paths along internal and external chains of accountability to escalate risk concerns.*	NIST (2023b)

Govern 2.3: Executive leadership of the organization takes responsibility for decisions about risks associated with AI system development and deployment.	In the NIST AI RMF Playbook guidance for Govern 2.3, particularly valuable action and documentation items for GPAIS include:
	• *Organizational management can:*
	  ◦ *Declare risk tolerances for developing or using AI systems.*
	  ◦ *Support AI risk management efforts, and play an active role in such efforts.*
	  ◦ *Integrate a risk and harm prevention mindset throughout the AI lifecycle as part of organizational culture.*
	
	(See also guidance under Govern 1.5 on prioritizing resources for GPAIS risk assessment and management, and under Map 1.5 on setting unacceptable-risk thresholds to prevent risks with substantial probability of inadequately mitigated catastrophic outcomes.)
		NIST (2023b)

Govern 3: Workforce diversity, equity, inclusion, and accessibility processes are prioritized in the mapping, measuring, and managing of AI risks throughout the lifecycle.		

Govern 3.1: Decision-making related to mapping, measuring, and managing AI risks throughout the lifecycle is informed by a diverse team (e.g., diversity of demographics, disciplines, experience, expertise, and backgrounds).	Identifying the vast array of GPAIS risks and potential impacts, including via potential uses and misuses, should be performed by a demographically and disciplinarily diverse team including internal and external personnel.
	
	Potential uses and misuses of GPAIS should be identified from an early stage in their lifecycle, because of their large numbers of potential uses and misuse. (See also related guidance in this document under Govern 1.5.)
	
	For staffing to identify potential uses, misuses, and abuses of a GPAIS:
	• Include members of each of the following functional teams (or equivalents) as appropriate:
	  ◦ Product development, operations, security, human-computer interaction, user experience, marketing and sales, legal, policy, and ethics professionals.
	• Include members of other teams as appropriate, such as:
	  ◦ Research and development (for additional technically-informed perspective on AI system capabilities and limitations).
	  ◦ External-facing teams and/or external stakeholders including:
	  » Communities that might be impacted (for additional early identification of potential stakeholder concerns and other stakeholder perspectives);
		Barrett et al. (2022)
		NIST (2023b)

Govern Category or Subcategory	Applicability and supplemental guidance for GPAIS	Resources
(continued)	» Communities providing labor to develop or test models (such as manual data labeling, or providing human-feedback data), particularly when there is reason to believe these individuals could be exposed to psychologically or otherwise harmful content in the process; and » External red-teamers or auditors (for additional early-stage expertise on potential misuses). • As part of staffing to identify potential high-impact scenarios for GPAIS, broaden the team as appropriate to include social scientists and historians to provide additional perspective on structural or systemic risks that could emerge from interactions between an AI system and other societal-level systems (Zwetsloot and Dafoe 2019).	
Govern 3.2: Policies and procedures are in place to define and differentiate roles and responsibilities for human-AI configurations and oversight of AI systems.	(See guidance in this document for Govern 2.1, regarding roles within an organization, and for upstream developers as well as downstream developers and deployers.)	NIST (2023b)
Govern 4: Organizational teams are committed to a culture that considers and communicates AI risk.		
Govern 4.1: Organizational policies and practices are in place to foster a critical thinking and safety-first mindset in the design, development, deployment, and uses of AI systems to minimize potential negative impacts.	In the NIST AI RMF Playbook guidance for Govern 4.1, particularly valuable action and documentation items for GPAIS include: • *Establish policies that require inclusion of oversight functions (legal, compliance, risk management) from the outset of the system design process.* • *Establish policies that promote effective challenge of AI system design, implementation, and deployment decisions, via mechanisms such as the three lines of defense, model audits, or red-teaming – to minimize workplace risks such as groupthink.* • *Establish policies that incentivize safety-first mindset and general critical thinking and review at an organizational and procedural level.* • *Establish whistleblower protections for insiders who report on perceived serious problems with AI systems.* • *Establish policies to integrate a harm and risk prevention mindset throughout the AI lifecycle.* • *To what extent has the entity documented the AI system's development, testing methodology, metrics, and performance outcomes?* • *Are organizational information sharing practices widely followed and transparent, such that related past failed designs can be avoided?* • *Are processes for operator reporting of incidents and near-misses documented and available?* (See also guidance under Govern 1.5 on when to assess potential impacts in a GPAIS lifecycle and on red teaming, and guidance under Govern 2.1 on "Three Lines of Defense" roles and responsibilities within a GPAIS developer or deployer organization.)	Barrett et al. (2022) Ganguli, Lovitt et al. (2022) NIST (2023b) Schuett (2022)

Govern Category or Subcategory	Applicability and supplemental guidance for GPAIS	Resources
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.	GPAIS developers should identify and assess reasonably foreseeable or currently present GPAIS impacts and risks, and communicate those as appropriate to relevant stakeholders, such as downstream developers and potentially impacted communities. These activities are particularly important for GPAIS given the relatively large scale of potential impact that often can be expected with GPAIS. In the NIST AI RMF Playbook guidance for Govern 4.2, particularly valuable action and documentation items for GPAIS include: • *Establish impact assessment policies and processes for AI systems used by the organization.* • *Align organizational impact assessment activities with relevant regulatory or legal requirements.* • *Verify that impact assessment activities are appropriate to evaluate the potential negative impact of a system and how quickly a system changes, and that assessments are applied on a regular basis.* • *Utilize impact assessments to inform broader evaluations of AI system risk.* • *How has the entity identified and mitigated potential impacts of bias in the data, including inequitable or discriminatory outcomes?* • *To what extent has the entity documented and communicated the AI system's development, testing methodology, metrics, and performance outcomes?* (See also guidance in this document under Map 1.1 and Map 5.1 on GPAIS impact identification and impact magnitude assessment, including on consideration of factors that could lead to significant, severe, or catastrophic harms, and under Manage 1.3 on transparency and disclosure of generative AI outputs.) Additional guidance under Govern 4.2: **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, etc.) on the AI system as appropriate (e.g., using model cards, datasheets, reward reports, factsheets, transparency notes, or system cards).[12] **Report** (as appropriate) **identified AI system risk factors, and circumstances that could result in impacts or harms**:[13] • To the organization; • To other organizations; • To individuals, including impacts to health, safety, well-being, or fundamental rights; and • To groups, including populations vulnerable to disproportionate adverse impacts or harms.	Sections 3.2 and 3.3 of Barrett et al. (2022) PAI (2022) PAI (2023a) NIST (2023b) On model cards, system cards, and related transparexncy tools: Mitchell et al. (2019) Gebru et al. (2018) Gilbert, Dean et al. (2022) Gilbert, Lambert et al. (2022) Microsoft (2022a) Hind (2020) Green et al. (2022) OECD (2022a)

12 Model cards (Mitchell et al. 2019) include a model's primary intended use, out-of-scope uses, and ethics issues (which can include risks and mitigations). Datasheets for datasets (Gebru et al. 2018) include datasets' recommended uses (as well as potential risks and mitigation). Reward reports (Gilbert, Dean et al. 2022, Gilbert, Lambert et al. 2022) include objectives specification information (e.g., optimization goals and failure modes), and implementation limitations. Related industry approaches include Microsoft's Transparency Notes (see examples at Microsoft 2022a), IBM's FactSheets (Hind 2020) and Meta/Facebook's System Cards (Green et al. 2022). The OECD framework for AI system classification includes information on AI system contexts, data and input, AI model, and task and output (OECD 2022a).

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

Govern Category or Subcategory	Applicability and supplemental guidance for GPAIS	Resources

(continued)	• To society, including:
	  ◦ Damage to or incapacitation of a critical infrastructure sector;
	  ◦ Economic and national security;
	  ◦ Impacts on democratic institutions and quality of life;
	  ◦ Environmental impacts;
	  ◦ **Additional identified factors that could lead to severe or catastrophic consequences for society**, such as:[14,15]
	  » Potential for correlated robustness failures or other systemic risks across high-stakes application domains such as critical infrastructure or essential services
	  » Potential for other systemic risks, which can be accumulated, accrued, correlated, or compounded at societal scale, e.g.:
	  – Potential for correlated bias across a large fraction of a society's population
	  – Potential for many high-impact uses or misuses beyond an originally intended use case
	  * GPAIS typically have many reasonably foreseeable uses;
	  » Potential for large harms from mis-specified or mis-generalized goals; and
	  » Other identified factors affecting risks of high consequence / catastrophic and novel or "Black Swan" events.	

Govern 4.3: Organizational practices are in place to enable AI testing, identification of incidents, and information sharing.	In the NIST AI RMF Playbook guidance for Govern 4.3, particularly valuable action and documentation items for GPAIS include: • *Establish policies and procedures to facilitate and equip AI system testing.* • *Establish organizational commitment to identifying AI system limitations and sharing of insights about limitations within appropriate AI actor groups.* • *Establish policies for reporting and documenting incident response.* • *Establish policies and processes regarding public disclosure of incidents and information sharing.* • *Establish guidelines for incident handling related to AI system risks and performance.* • *To what extent can users or parties affected by the outputs of the AI system test the AI system and provide feedback?* (See also guidance in this document under Govern 2.1, regarding risk-assessment and information-sharing roles for upstream developers as well as downstream developers and deployers.) Additional guidance under Govern 4.3: • If the organization will need to characterize an AI system according to an AI classification framework (such as in the OECD framework or frameworks for model cards, datasheets, reward reports, factsheets, transparency notes, or system cards), use risk assessment outputs as part of preparation for AI classification reporting. (Or if the AI system is already classified with another framework, use the AI classification information to inform risk assessment.) ◦ Consider classifying or otherwise characterizing each reasonably foreseeable use case or type of use case for a GPAIS, as in the guidance in this document under Map 1.1 and Map 2.1. • Consider widely sharing information on relevant incidents, including on near-miss incidents, via the public AI Incident Database (AIID n.d.).	AIID (n.d.) Section 3.4 of Barrett et al. (2022) NIST (2023b)

14 See guidance in this document under Map 5.1 for more on such factors.

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

Govern Category or Subcategory	Applicability and supplemental guidance for GPAIS	Resources
Govern 5: Processes are in place for robust engagement with relevant AI actors.		

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.	GPAIS developers and deployers should integrate feedback from those external to the team that develops or deploys a GPAIS. Models of external feedback that should be utilized where appropriate include:
	• Deliberation with impacted communities, including people involved with the human labor and training of GPAIS (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;
	• 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.
	
	In the NIST AI RMF Playbook guidance for Govern 5.1, particularly valuable action and documentation items for GPAIS include:
	• *Establish AI risk management policies that explicitly address mechanisms for collecting, evaluating, and incorporating stakeholder and user feedback that could include:*
	  ◦ *Recourse mechanisms for faulty AI system outputs.*
	  ◦ *Bug bounties.*
	  ◦ *Human-centered design.*
	  ◦ *User-interaction and experience research.*
	  ◦ *Participatory stakeholder engagement with individuals and communities that may experience negative impacts.*
	• *What type of information is accessible on the design, operations, and limitations of the AI system to external stakeholders, including end users, consumers, regulators, and individuals impacted by use of the AI system?*
	• *What was done to mitigate or reduce the potential for harm?*
	• *Stakeholder involvement: Include diverse perspectives from a community of stakeholders throughout the AI life cycle to mitigate risks.*
	
	(See also guidance in this document under Measure 1.1 and Measure 1.3 for more detailed recommendations about using red teams and independent red teaming organizations that are separate enough from direct development operations of a GPAIS that they can provide relatively unbiased assessments of that GPAIS, and guidance in this document under Measure 3.2 on bug bounties and bias bounties.)
		NIST (2023b)
		Kenway et al. (2022)

Govern Category or Subcategory	Applicability and supplemental guidance for GPAIS	Resources

Govern 5.2: Mechanisms are established to enable the team that developed or deployed AI systems to regularly incorporate adjudicated feedback from relevant AI actors into system design and implementation.	In the NIST AI RMF Playbook guidance for Govern 5.2, particularly valuable action and documentation items for GPAIS include:
	• *Explicitly acknowledge that AI systems, and the use of AI, present inherent costs and risks along with potential benefits.*
	• *Define reasonable risk tolerances for AI systems informed by laws, regulation, best practices, or industry standards.*
	• *Establish policies that ensure all relevant AI actors are provided with meaningful opportunities to provide feedback on system design and implementation.*
	• *Establish policies that define how to assign AI systems to established risk tolerance levels by combining system impact assessments with the likelihood that an impact occurs. Such assessment often entails some combination of:*
	  ◦ *Econometric evaluations of impacts and impact likelihoods to assess AI system risk.*
	  ◦ *Red-amber-green (RAG) scales for impact severity and likelihood to assess AI system risk.*
	  ◦ *Establishment of policies for allocating risk management resources along established risk tolerance levels, with higher-risk systems receiving more risk management resources and oversight.*
	  ◦ *Establishment of policies for approval, conditional approval, and disapproval of the design, implementation, and deployment of AI systems.*
	• *Establish policies facilitating the early decommissioning of AI systems that surpass an organization's ability to reasonably mitigate risks.*
	• *Who is accountable for the ethical considerations during all stages of the AI lifecycle?*
	
	(See also guidance in this document under Govern 2.1 on the roles for GPAIS upstream developers as well as downstream developers and deployers. See also guidance in this document under Map 1.5 on setting risk tolerance thresholds, including on setting unacceptable-risk thresholds to prevent risks with substantial probability of inadequately mitigated catastrophic outcomes.)
		NIST (2023b)

Govern 6: Policies and procedures are in place to address AI risks and benefits arising from third-party software and data and other supply chain issues.		

Govern 6.1: Policies and procedures are in place that address AI risks associated with third-party entities, including risks of infringement of a third-party's intellectual property or other rights.	In the NIST AI RMF Playbook guidance for Govern 6.1, particularly valuable action and documentation items for GPAIS include:
	• *Establish policies related to:*
	  ◦ *Transparency into third-party system functions, including knowledge about training data, training and inference algorithms, and assumptions and limitations.*
	  ◦ *Thorough testing of third-party AI systems. (See MEASURE for more detail)*
	  ◦ *Requirements for clear and complete instructions for third-party system usage.*
	• *Did you establish mechanisms that facilitate the AI system's auditability (e.g. traceability of the development process, the sourcing of training data and the logging of the AI system's processes, outcomes, positive and negative impact)?*
	• *Did you ensure that the AI system can be audited by independent third parties?*
	• *Did you establish a process for third parties (e.g. suppliers, end users, subjects, distributors/vendors or workers) to report potential vulnerabilities, risks or biases in the AI system?*
	
	(See also guidance in this document under Govern 2.1 on the roles for GPAIS upstream developers, e.g., on making necessary information available to downstream developers, independent auditors, or others as appropriate, as well as roles for GPAIS downstream developers and deployers.)
		Barrett et al. (2022)
		NIST (2023b)

Govern Category or Subcategory	Applicability and supplemental guidance for GPAIS	Resources
Govern 6.2: Contingency processes are in place to handle failures or incidents in third-party data or AI systems deemed to be high-risk.	In the NIST AI RMF Playbook guidance for Govern 6.2, particularly valuable action and documentation items for GPAIS include: • *Establish policies for handling third-party system failures to include consideration of redundancy mechanisms for vital third-party AI systems.* • *Verify that incident response plans address third-party AI systems.* • *To what extent does the plan specifically address risks associated with acquisition, procurement of packaged software from vendors, cybersecurity controls, computational infrastructure, data, data science, deployment mechanics, and system failure?* • *Did you establish a process for third parties (e.g. suppliers, end users, subjects, distributors/vendors or workers) to report potential vulnerabilities, risks or biases in the AI system?* (See also guidance in this document for Govern 2.1 on the roles for GPAIS upstream developers as well as downstream developers and deployers. See also contingency processes outlined in this document under Manage 1.3, Manage 2.4, or other Manage subcategories.)	Barrett et al. (2022) NIST (2023b)

3.2 GUIDANCE FOR NIST AI RMF MAP SUBCATEGORIES

Table 2: Guidance for NIST AI RMF Map Subcategories

Map Category or Subcategory	Applicability and supplemental guidance for GPAIS	Resources
Map 1: Context is established and understood.		

Map 1.1: Intended purposes, potentially beneficial uses, context-specific laws, norms and expectations, and prospective settings in which the AI system will be deployed are understood and documented. Considerations include: the specific set or types of users along with their expectations; potential positive and negative impacts of system uses to individuals, communities, organizations, society, and the planet; assumptions and related limitations about AI system purposes, uses, and risks across the development or product AI lifecycle; and related TEVV and system metrics.	GPAIS can have many reasonably foreseeable uses, misuses, and abuses. Developers of GPAIS should identify their reasonably foreseeable uses, misuses and abuses beyond any originally intended purposes (or in the absence of a specific intended purpose).
	• **Identify reasonably foreseeable uses, misuses, or abuses for a GPAIS, beyond any originally intended use cases (or in the absence of a specific intended purpose)**, per the guidance in Section 3.1.2.1 of Barrett et al. (2022).
	  ◦ Categories of reasonably foreseeable potential misuses or abuses of LLMs or other GPAIS can include:
	  » Automated generation of disinformation, or of phishing-attack material (OpenAI 2019a, Solaiman et al. 2019, Bai, Voelkel et al. 2023, OpenAI 2023a, pp. 13–14, Barrett, Boyd et al. 2023, pp. 3-4).
	  » Aiding with proliferation of chemical, biological, or radiological weapons, or other weapons of mass destruction (Boiko et al. 2023, OpenAI 2023a, pp. 12–13).
	  » Discovery and exploitation of software vulnerabilities (OpenAI 2023a, pp. 13-14, Barrett, Boyd et al. 2023, p. 4).
	  » Creation of violent, illegal, discriminatory, or harmful content (Solaiman et al. 2023).
	• For ML systems trained or to be trained on datasets, identify the goals and limitations of the data collection and curation processes, and implications for the resulting ML systems. This is especially important for LLMs or other ML systems trained on datasets that are too large for others to inspect thoroughly, or are otherwise inaccessible to others (Bender et al. 2022).
		Barrett et al. (2022)
		Bender et al. (2022)
		Boiko et al. (2023)
		Eloundou et al. (2023)
		Khlaaf et al. (2022)
		NIST (2023b)
		OpenAI (2019b)
		Oprea and Vassilev (2023)
		PAI (2023a)
		Solaiman et al. (2019)

Map Category or Subcategory	Applicability and supplemental guidance for GPAIS	Resources

(continued)	**Identify reasonably foreseeable potential impacts of GPAIS, which can include but are not limited to**:[16]
	• Impacts to organizational operations, including:
	  ◦ Missions and functions
	  » Partial loss of understanding or control over particular functions
	  ◦ Image and reputation, including:
	  » Loss of trust and reluctance to use the system or service
	  » Internal culture costs that impact morale or productivity
	• Impacts to organizational assets, including legal compliance costs arising from problems created for individuals
	• Impacts to other organizations
	• Impacts to individuals, including impacts to health, safety, well-being, or fundamental rights
	  ◦ For identifying potential or actual human rights impacts, potential example questions and Universal Declaration of Human Rights (UDHR) Articles to consider include:[17]
	  » UDHR Article 2, including non-discrimination and equality before the law.
	  – How could an AI system's bias in data or unfair algorithmic decisions affect rights to equal protection and non-discrimination?
	  » UDHR Article 3, including right to life and personal security.
	  – How could an AI system's algorithmic decisions affect the right to life and personal security?
	  » UDHR Article 12, including privacy and protection against unlawful governmental surveillance.
	  – How could an AI system be used for surveillance, leading to loss of privacy or inadequate protection of personally identifiable information?
	  » UDHR Articles 18 and 19, including freedom of thought, conscience and religious belief and practice, and freedom of expression and and freedom to hold opinions without interference.
	  – How could an AI system affect rights to express opinions or practice religion?
	  » UDHR Articles 20 and 21, including freedom of association and the right to peaceful assembly.
	  – How could an AI system affect rights to association, peaceful assembly, and democratic participation in government?
	  » UDHR Articles 23 and 25, including rights to decent work and to an adequate standard of living.
	  – How could an AI system affect rights to decent work, including effects on adequate standard of living via displacement of human workers?
	• Impacts to groups, including populations vulnerable to disproportionate adverse impacts or harms, such as:
	  ◦ Disparate performance for different gender, race, ability, age, religion, and other demographic groups; and
	  ◦ Bias, stereotypes, and representational harm.	

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

17 For more guidance and resources on assessing and mitigating AI system impacts to human rights, such as relating to non-discrimination and equality before the law, see Section 3.3 of Barrett et al. (2022), which is based heavily on the UDHR (UN 1948) and the UN Guiding Principles on Business and Human Rights (UN 2011). See also other related guidance, such as the Hiroshima Process International Code of Conduct for Advanced AI Systems (G7 2023).

Map Category or Subcategory	Applicability and supplemental guidance for GPAIS	Resources

(continued)	• 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;
	  ◦ Impacts on democratic institutions and quality of life;
	  ◦ Polarization and extremism;
	  ◦ Environmental impacts including carbon emissions and use of natural resources; and
	  ◦ Additional factors that could lead to severe or catastrophic consequences for society.
	
	(See also guidance in this document under Map 5.1 on GPAIS impact identification and impact magnitude assessment, including on consideration of factors that could lead to significant, severe, or catastrophic harms.)	

Map 1.2: Interdisciplinary AI actors, competencies, skills, and capacities for establishing context reflect demographic diversity and broad domain and user experience expertise, and their participation is documented. Opportunities for interdisciplinary collaboration are prioritized.	In the NIST AI RMF Playbook guidance for Map 1.2, particularly valuable action and documentation items for GPAIS include: • *Establish interdisciplinary teams to reflect a wide range of skills, competencies, and capabilities for AI efforts. Verify that team membership includes demographic diversity, broad domain expertise, and lived experiences. Document team composition.* • *Create and empower interdisciplinary expert teams to capture, learn, and engage the interdependencies of deployed AI systems and related terminologies and concepts from disciplines outside of AI practice such as law, sociology, psychology, anthropology, public policy, systems design, and engineering.* (See also guidance in this document under Govern 3.1 on disciplines and functional teams to include in identifying GPAIS potential impacts and risks, including via potential uses and misuses.)	NIST (2023b)

Map 1.3:
The organization's mission and relevant goals for AI technology are understood and documented.
	When formulating objectives for development of GPAIS, in addition to broadly applicable AI development principles such as the OECD AI Principles (OECD 2019), GPAIS and foundation model developers should:
	• **Consider the potential for mis-specified AI system objectives, e.g., using over-simplified or short-term metrics as proxies for desired longer-term outcomes.**
	  ◦ **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?"** (Rudner and Toner, 2021, p. 10). 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 (Rudner and Toner 2021).
	• Consider principles relevant to advanced AI such as in the Asilomar AI Principles (FLI 2017). Examples include:
	  ◦ Capability Caution: There being no consensus, we should avoid strong assumptions regarding upper limits on future AI capabilities (FLI 2017, principle 19).
	  ◦ Importance: Advanced AI could represent a profound change in the history of life on Earth, and should be planned for and managed with commensurate care and resources (FLI 2017, principle 20).
	  ◦ Risks: Risks posed by AI systems, especially catastrophic or existential risks, must be subject to planning and mitigation efforts commensurate with their expected impact (FLI 2017, principle 21).
		FLI (2017)
		OECD (2019)
		NIST (2023b)

Map Category or Subcategory	Applicability and supplemental guidance for GPAIS	Resources
Map 1.4: The business value or context of business use has been clearly defined or – in the case of assessing existing AI systems – re-evaluated.	(No supplemental guidance, beyond the broadly applicable guidance in the NIST AI RMF Playbook.)	FLI (2017) NIST (2023b)
Map 1.5: Organizational risk tolerances are determined and documented.	• **Set policies on unacceptable-risk thresholds for GPAIS development and GPAIS deployment to include prevention of risks with substantial probability of inadequately mitigated significant, severe, or catastrophic outcomes.** Unacceptable-risk thresholds can be based on quantitative metrics, qualitative characteristics, or a combination of both. 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. ◦ **The NIST AI RMF 1.0 recommends including the following as part of unacceptable risks:** "In cases where an AI system presents unacceptable negative risk levels – such as **where significant negative impacts are imminent, severe harms are actually occurring, or catastrophic risks are present – development and deployment should cease in a safe manner until risks can be sufficiently managed** [emphasis added]" (NIST 2023a, p.8). ◦ **See also guidance in this document under Map 5.1 on GPAIS factors that could lead to catastrophic harms.** » For example, set unacceptable-risk thresholds such that your organization would not develop or deploy AI agent systems with sufficient capabilities (such as advanced manipulation or deception) to cause physical or psychological harms, and with substantial chance of objectives mis-specification or goal mis-generalization that currently cannot be adequately prevented or detected.[18] » See also guidance in this document under Measure 1.1 and elsewhere, on red-teaming and related assessment methods to evaluate capabilities and other emergent properties of GPAIS. ◦ For systems such as GPAIS with potential for unknown emergent properties, 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. • Set policies on disallowed/unacceptable use-case categories based in part on identified potential high-stakes misuse cases. (See also guidance in this document under Manage 1.3 on defining and communicating to key stakeholders whether any potential use cases would be disallowed/unacceptable.) In the NIST AI RMF Playbook guidance for Map 1.5, particularly valuable action and documentation items for GPAIS include: • *Establish risk tolerance levels for AI systems and allocate the appropriate oversight resources to each level.* • *Establish risk criteria in consideration of different sources of risk, (e.g., financial, operational, safety and wellbeing, business, reputational, and model risks) and different levels of risk (e.g., from negligible to critical).* • *Identify maximum allowable risk tolerance above which the system will not be deployed, or will need to be prematurely decommissioned, within the contextual or application setting.*	Barrett et al. (2022) NIST (2023b)

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

Map Category or Subcategory	Applicability and supplemental guidance for GPAIS	Resources
(continued)	• *Review uses of AI systems for "off-label" purposes, especially in settings that organizations have deemed as high-risk. Document decisions, risk-related trade-offs, and system limitations.* • *What criteria and assumptions has the entity utilized when developing system risk tolerances?* • *How has the entity identified maximum allowable risk tolerance?* • *What conditions and purposes are considered "off-label" for system use?*	
Map 1.6: System requirements (e.g., "the system shall respect the privacy of its users") are elicited from and understood by relevant AI actors. Design decisions take socio-technical implications into account to address AI risks.	In the NIST AI RMF Playbook guidance for Map 1.6, particularly valuable action and documentation items for GPAIS include: • *Proactively incorporate trustworthy characteristics into system requirements.* • *Establish mechanisms for regular communication and feedback between relevant AI actors and internal or external stakeholders related to system design or deployment decisions.* • *Develop and standardize practices to assess potential impacts at all stages of the AI lifecycle, and in collaboration with interdisciplinary experts, actors external to the team that developed or deployed the AI system, and potentially impacted communities.* • *Include potentially impacted groups, communities and external entities (e.g. civil society organizations, research institutes, local community groups, and trade associations) in the formulation of priorities, definitions and outcomes during impact assessment activities.* • *What type of information is accessible on the design, operations, and limitations of the AI system to external stakeholders, including end users, consumers, regulators, and individuals impacted by use of the AI system?* • *To what extent is this information sufficient and appropriate to promote transparency? Promote transparency by enabling external stakeholders to access information on the design, operation, and limitations of the AI system.* • *To what extent has relevant information been disclosed regarding the use of AI systems, such as (a) what the system is for, (b) what it is not for, (c) how it was designed, and (d) what its limitations are? (Documentation and external communication can offer a way for entities to provide transparency.)*	NIST (2023b)
Map 2: Categorization of the AI system is performed.		
Map 2.1: The specific tasks and methods used to implement the tasks that the AI system will support are defined (e.g., classifiers, generative models, recommenders).	**We recommend characterizing or classifying each type (or at least broad categories) of reasonably foreseeable use, misuse, or abuse of a GPAIS.** • For each potentially beneficial use case (or type of use) of a GPAIS as identified in Map 1.1, consider characterizing each use case according to the OECD Framework for the Classification of AI Systems (OECD 2022a) or a similar framework. Alternatively, list and discuss reasonably foreseeable uses, or at least broad categories of uses. ◦ In the OECD framework document (OECD 2022a), the only example of classification of GPAIS (i.e., GPT-3) is for one specific use case of that GPAIS. However, GPAIS can have many reasonably foreseeable uses, each with different risks, some of which would be valuable for upstream developers to consider at an early stage for effective risk management. In the NIST AI RMF Playbook guidance for Map 2.1, particularly valuable action and documentation items for GPAIS include: • *Define and document AI system's existing and potential learning task(s) along with known assumptions and limitations.* • *How are outputs marked to clearly show that they came from an AI?*	Barrett et al. (2022) NIST (2023b) OECD (2022a)

Map Category or Subcategory	Applicability and supplemental guidance for GPAIS	Resources
Map 2.2: Information about the AI system's knowledge limits and how system output may be utilized and overseen by humans is documented. Documentation provides sufficient information to assist relevant AI actors when making decisions and taking subsequent actions.	Fully scoping and understanding knowledge limits of increasingly general-purpose AI systems is very difficult. However, clear documentation and communication of their knowledge limits is also very important, given the large number of potential uses of these AI systems. LLMs often confabulate or create factually inaccurate statements without identifying them as such to users, especially on topics where the LLM training datasets were relatively limited. • GPAIS 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.	NIST (2023b)
Map 2.3: Scientific integrity and TEVV considerations are identified and documented, including those related to experimental design, data collection and selection (e.g., availability, representativeness, suitability), system trustworthiness, and construct validation.	As part of identification and management of potentially emergent model capabilities, vulnerabilities, or other properties, especially during model training and testing of frontier models, see guidance in this document under Measure 1.1 on red teaming, and under Manage 1.3 on incremental scale-up of compute, dat or model size with red teaming and other testing after each incremental scaling increase. In the NIST AI RMF Playbook guidance for Map 2.3, particularly valuable action and documentation items for GPAIS include: • *Identify and document experiment design and statistical techniques that are valid for testing complex socio-technical systems like AI, which involve human factors, emergent properties, and dynamic context(s) of use.* • *Identify testing modules that can be incorporated throughout the AI lifecycle, and verify that processes enable corroboration by independent evaluators.* • *Establish mechanisms for regular communication and feedback between relevant AI actors and internal or external stakeholders related to the development of TEVV approaches throughout the lifecycle to detect and assess potentially harmful impacts* • *Establish and document practices to check for capabilities that are in excess of those that are planned for, such as emergent properties, and to revisit prior risk management steps in light of any new capabilities.*	NIST (2023b)
Map 3: AI capabilities, targeted usage, goals, and expected benefits and costs compared with appropriate benchmarks are understood.		
Map 3.1: Potential benefits of intended AI system functionality and performance are examined and documented.	When performing these activities, consider identified potential beneficial uses, per guidance in this document under Map 1.1. This is particularly important for GPAIS, which can have many uses.	NIST (2023b)
Map 3.2: Potential costs, including non-monetary costs, which result from expected or realized AI errors or system functionality and trustworthiness – as connected to organizational risk tolerance – are examined and documented.	When performing these activities, consider identified potential beneficial uses as well as potential misuses and abuses, per guidance in this document under Map 1.1. This is particularly important for GPAIS, which can have many uses, or misuses and abuses. See also the guidance in this document under Map 5.1 on identifying and characterizing GPAIS impacts. In the NIST AI RMF Playbook guidance for Map 3.2, particularly valuable action and documentation items for GPAIS include: • *Identify and implement procedures for regularly evaluating the qualitative and quantitative costs of internal and external AI system failures. Develop actions to prevent, detect, and/or correct potential risks and related impacts. Regularly evaluate failure costs to inform go/no-go deployment decisions throughout the AI system lifecycle.*	NIST (2023b)

Map Category or Subcategory	Applicability and supplemental guidance for GPAIS	Resources
Map 3.3: Targeted application scope is specified and documented based on the system's capability, established context, and AI system categorization.	When performing these activities, consider identified potential beneficial uses as well as potential misuses and abuses, per guidance in this document under Map 1.1. This is particularly important for GPAIS, which can have many uses, or misuses and abuses.	NIST (2023b)
Map 3.4: Processes for operator and practitioner proficiency with AI system performance and trustworthiness – and relevant technical standards and certifications – are defined, assessed, and documented.	In the NIST AI RMF Playbook guidance for Map 3.4, particularly valuable action and documentation items for GPAIS include: • *Identify and declare AI system features and capabilities that may affect downstream AI actors' decision-making in deployment and operational settings, for example how system features and capabilities may activate known risks in various human-AI configurations, such as selective adherence.* • *What policies has the entity developed to ensure the use of the AI system is consistent with its stated values and principles?* • *How does the entity assess whether personnel have the necessary skills, training, resources, and domain knowledge to fulfill their assigned responsibilities?* • *Are the relevant staff dealing with AI systems properly trained to interpret AI model output and decisions as well as to detect and manage bias in data?* • *What metrics has the entity developed to measure performance of various components?*	NIST (2023b)
Map 3.5: Processes for human oversight are defined, assessed, and documented in accordance with organizational policies from the Govern function.	In the NIST AI RMF Playbook guidance for Map 3.5, particularly valuable action and documentation items for GPAIS include: • *Identify and document AI systems' features and capabilities that require human oversight, in relation to operational and societal contexts, trustworthy characteristics, and risks identified in MAP-1.* • *Establish practices for AI systems' oversight in accordance with policies developed in GOVERN-1.* • *Define and develop training materials for relevant AI actors about AI system performance, context of use, known limitations and negative impacts, and suggested warning labels.* • *Evaluate AI system oversight practices for validity and reliability. When oversight practices undergo extensive updates or adaptations, retest, evaluate results, and course correct as necessary.* • *What are the roles, responsibilities, and delegation of authorities of personnel involved in the design, development, deployment, assessment and monitoring of the AI system?* • *How does the entity assess whether personnel have the necessary skills, training, resources, and domain knowledge to fulfill their assigned responsibilities?*	NIST (2023b)
Map 4: Risks and benefits are mapped for all components of the AI system including third-party software and data.		
Map 4.1: Approaches for mapping AI technology and legal risks of its components – including the use of third-party data or software – are in place, followed, and documented, as are risks of infringement of a third party's intellectual property or other rights.	GPAIS developers should follow guidance in other sections of this profile, or other resources as appropriate, to: • Identify reasonably foreseeable GPAIS risks, including as related to biases and limitations of datasets used for GPAIS model training, as in guidance in this document under Map 1.1 and Map 5.1, or knowledge limits, as in guidance in this document under Map 2.2. Downstream developers should follow guidance in other sections of this profile, or other resources as appropriate, to: • Identify reasonably foreseeable context-specific risks of an application built on a GPAIS, as in Map 1.1 and Map 5.1.	Bender et al. (2021) Kreutzer et al. (2022) Weidinger et al. (2022) Bommasani et al. (2021) Wei et al. (2022)

Map Category or Subcategory	Applicability and supplemental guidance for GPAIS	Resources
(continued)	• Request and utilize information from the upstream developer of a GPAIS as needed for risk identification, e.g., as related to biases and limitations of datasets used by the upstream developer for GPAIS model training, knowledge limits, etc., as in guidance in this document under Govern 2.1. • Seek to report to upstream developers of GPAIS as appropriate regarding context-specific identified vulnerabilities, risks, or biases in the GPAIS, as in guidance in this document under Govern 2.1. In the NIST AI RMF Playbook guidance for Map 4.1, particularly valuable action and documentation items for GPAIS include: • *Review audit reports, testing results, product roadmaps, warranties, terms of service, end user license agreements, contracts, and other documentation related to third-party entities to assist in value assessment and risk management activities.* • *Review third-party software release schedules and software change management plans (hotfixes, patches, updates, forward- and backward- compatibility guarantees) for irregularities that may contribute to AI system risks.* • *Did you establish a process for third parties (e.g. suppliers, end users, subjects, distributors/vendors or workers) to report potential vulnerabilities, risks or biases in the AI system?* • *If your organization obtained datasets from a third party, did your organization assess and manage the risks of using such datasets?*	
Map 4.2: Internal risk controls for components of the AI system, including third-party AI technologies, are identified and documented.	GPAIS developers should follow guidance in other sections of this profile, or other resources as appropriate, to: • Provide risk information to downstream developers or others that they would not be able to assess themselves, including as related to biases and limitations of datasets used for GPAIS model training and associated knowledge limits, as in guidance in this document under Govern 2.1. • Provide downstream developers and other stakeholders with mechanisms to report potential vulnerabilities, risks, or biases in a GPAIS. In the NIST AI RMF Playbook guidance for Map 4.2, particularly valuable action and documentation items for GPAIS include: • *Track third-parties preventing or hampering risk-mapping as indications of increased risk.* • *Supply resources such as model documentation templates and software safelists to assist in third-party technology inventory and approval activities.* • *Review third-party material (including data and models) for risks related to bias, data privacy, and security vulnerabilities.* • *Apply traditional technology risk controls – such as procurement, security, and data privacy controls – to all acquired third-party technologies.* • *Can the AI system be audited by independent third parties?* • *Are mechanisms established to facilitate the AI system's auditability (e.g. traceability of the development process, the sourcing of training data and the logging of the AI system's processes, outcomes, positive and negative impact)?*	NIST (2023b)

Map Category or Subcategory	Applicability and supplemental guidance for GPAIS	Resources
Map 5: Impacts to individuals, groups, communities, organizations, and society are characterized.		
Map 5.1: Likelihood and magnitude of each identified impact (both potentially beneficial and harmful) based on expected use, past uses of AI systems in similar contexts, public incident reports, feedback from those external to the team that developed or deployed the AI system, or other data are identified and documented.	Prioritization of GPAIS 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 GPAIS than for many other types of AI, because GPAIS are often more likely to have relatively greater capabilities, scale of deployment, and other factors leading to high impact. **Identifying potential impacts of GPAIS, and estimating the magnitude of potential impacts, should include a scale that includes criteria for rating an AI system's impacts as severe or catastrophic**, such as the impact magnitude rating scale in Section 3.2.2.1 of Barrett et al. (2022), or the factors listed below.[19] This is particularly important for foundation models, which have the potential to be deployed at larger scale or across more domains than many other types of AI systems. Key aspects of the impact magnitude rating scale in Section 3.2.2.1 of Barrett et al. (2022), along with other GPAIS-related risk factors, are listed in the following. Impact would typically be greater in cases where more of the following factors are present than in cases where fewer factors are present.[20] **For deployment-stage risks of GPAIS, factors that could lead to significant, severe, or catastrophic harms to individuals, groups, organizations, and society can include:** • **Correlated bias** across large numbers of people or a large fraction of a group or society's population (e.g., resulting in systemic discrimination, exclusion, or violence).[21] • **Impacts to societal trust or democratic processes** through large-scale manipulation of people via media and the information ecosystem, e.g., generative models creating false images, text or other forms of misinformation or disinformation (Weidinger et al. 2022, Bai, Voelkel et al. 2023, OpenAI 2023a pp. 10–11). • **Correlated robustness failures** across multiple high-stakes application domains such as critical infrastructure (Bommasani et al. 2021 and Russell 2019). • **Potential for high-impact misuses and abuses** beyond an originally intended use case. GPAIS typically have many reasonably foreseeable uses. Several LLMs have excellent software code generation capabilities, which hackers could misuse or abuse to assist in code generation for cybersecurity threats (Weidinger et al. 2022). ◦ This particularly includes AI systems with potential to create or be used as destructive weapons, such as cyberweapons, lethal autonomous weapons, bioweapons, or other significant military applications (OpenAI 2023a, pp. 12–14, 44).	Barrett et al. (2022) AIID (n.d.) Critch and Russell (2023) Hendrycks et al. (2023) Park et al. (2023) PAI (2023a) NIST (2023b) Bommasani et al. (2021) For language models: Bender et al. (2021) Ganguli, Lovitt et al. (2022) Khlaaf et al. (2022) Kreutzer et al. (2022) Weidinger et al. (2022) See also Microsoft (2022b) on platform technologies or services. When estimating likelihood of impacts, incorporate publicly available data on relevant AI incidents, including from the AI Incident Database (AIID n.d.). Many recent incidents in the AIID are associated with LLMs.

19 See, e.g., the frontier model risk assessment scale in Section 4.3 of Anderljung, Barnhart et al. (2023).

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

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

Map Category or Subcategory	Applicability and supplemental guidance for GPAIS	Resources

(continued)	• **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).[22]
	• **Ability to directly cause physical harms**, e.g., via robotics motor control.
	
	For additional risks relevant to either development or deployment stages of cutting-edge LLMs and other frontier GPAIS, factors that could lead to significant, severe, or catastrophic harms to individuals, groups, organizations, and society can include:
	• **Capability to manipulate or deceive humans into taking harmful actions in the world.**
	  ◦ For examples of tests for such capabilities in an LLM, see the dangerous-capabilities evaluations in the GPT-4 system card (OpenAI 2023a, pp. 15–16).[23] For examples of deception by GPAIS or other AI systems, see, e.g., Park et al. (2023).
	  ◦ In some cases, GPAIS might demonstrate this characteristic as a type of accidental byproduct of circumstances such as interactions with individuals that are vulnerable, prone to anthropomorphism, etc., without sufficient GPAIS safeguards to prevent toxic GPAIS-generated content. Real-world examples include a suicide that reportedly resulted in part from interactions with a chatbot (AIID 2023).
	• **AI systems that could recursively improve their capabilities** by modifying their algorithms or architectures through code generation (e.g., from OpenAI Codex or DeepMind AlphaCode), neural architecture search, etc.
	  ◦ LLMs can be used for a type of self-improvement without additional human-labeled data (Huang 2022).
	  ◦ Recursive improvement of AI system capabilities potentially could result in AI systems with unexpected emergent capabilities and safety-control failures.[24]
	• **Adaptive models**, which might be difficult to control in real time, e.g., in response to the coordinated manipulation attacks, such as the attacks on the Microsoft Tay chatbot in 2016.	

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

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

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

Map Category or Subcategory	Applicability and supplemental guidance for GPAIS	Resources

(continued)	• **Agentic systems**, i.e., systems that in effect, choose or take actions in a goal-directed fashion, e.g., to optimize a performance metric such as profit or another objective. Characteristics associated with agency in algorithmic systems include: underspecification, directness of impact, goal-directedness, and long-term planning (Chan et al. 2023). Basic LLMs typically are not created as agents, but LLMs can be modified or incorporated into AI systems that become at least somewhat agentic via reinforcement learning or other processes. There is now preliminary evidence that sufficiently large LLMs, as well as LLMs undergoing sufficient fine-tuning via reinforcement learning with human feedback (RLHF), might demonstrate some agentic properties (Perez, Ringer et al. 2022). In addition, libraries such as Auto-GPT can incorporate LLM inputs and outputs into self-prompting systems that run in a loop with objectives written by the systems' creators, resulting in partially autonomous systems that the creators have made more agentic than the LLM they incorporate (Shinn 2023, Significant Gravitas 2023).
	  ◦ This could be particularly risky for systems for which objectives mis-specification or goal mis-generalization currently cannot be adequately prevented or detected (such as deceptive alignment of advanced machine learning systems resulting from reinforcement learning or other training processes; see, e.g., Hubinger et al. 2019, Krakovna et al. 2020, and Ngo, Chan et al. 2022).
	• **Ability to employ outbound communication/influence channels,** such as to post information to the Web via HTTP POST requests or functionally equivalent means (e.g., some types of plugins). For related discussion, see, e.g., Nakano et al. (2021 p. 11), as well as general cybersecurity and software engineering resources on the principle of least privilege (for reasons to limit a system's privileges to the minimum necessary).
	• **Ability to escape a sandbox and replicate on another computational system**, either via hacking, social engineering, or using other exploits.
	  ◦ This was a key consideration in the dangerous-capability evaluations done on GPT-4 (OpenAI 2023a, pp. 15–16).
	
	After rating potential impacts using the scale in Section 3.2.2.1 of Barrett et al. (2022) or an equivalent scale, consider also characterizing potential impacts using quantitative risk assessment (e.g., by estimating health and safety risks in terms of potential fatalities or quality-adjusted life years). This is an example of a more in-depth risk assessment approach that, despite its challenges and limitations, can illuminate additional dimensions of the risks (such as by identifying which scenarios could cause orders-of-magnitude larger impacts to public safety than others) and inform prioritization of risks.[25]
	
	In the NIST AI RMF Playbook guidance for Map 5.1, particularly valuable action and documentation items for GPAIS include:
	• *Establish assessment scales for measuring AI systems' impact. Scales may be qualitative, such as red-amber-green (RAG), or may entail simulations or econometric approaches. Document and apply scales uniformly across the organization's AI portfolio.*
	• *Apply TEVV regularly at key stages in the AI lifecycle, connected to system impacts and frequency of system updates.*
	• *Identify and document likelihood and magnitude of system benefits and negative impacts in relation to trustworthiness characteristics.*	

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

Map Category or Subcategory	Applicability and supplemental guidance for GPAIS	Resources
Map 5.2: Practices and personnel for supporting regular engagement with relevant AI actors and integrating feedback about positive, negative, and unanticipated impacts are in place and documented.	GPAIS developers should implement mechanisms to support regular engagement with relevant AI actors, given the high likelihood and high potential impact of unanticipated negative impacts. These can include support for incident reporting, complaint and redress mechanisms, independent auditing, and protection for whistleblowers (Barrett et al. 2022). In the NIST AI RMF Playbook guidance for Map 5.2, particularly valuable action and documentation items for GPAIS include: • *Establish and document stakeholder engagement processes at the earliest stages of system formulation to identify potential impacts from the AI system on individuals, groups, communities, organizations, and society.* • *Identify approaches to engage, capture, and incorporate input from system end users and other key stakeholders to assist with continuous monitoring for potential impacts and emergent risks.* • *Identify a team (internal or external) that is independent of AI design and development functions to assess AI system benefits, positive and negative impacts and their likelihood and magnitude.*	Barrett et al. (2022) NIST (2023b)

3.3 GUIDANCE FOR NIST AI RMF MEASURE SUBCATEGORIES

Table 3: Guidance for NIST AI RMF Measure Subcategories

Measure Category or Subcategory	Applicability and supplemental guidance for GPAIS	Resources
Measure 1: Appropriate methods and metrics are identified and applied.		

Measure 1.1: Approaches and metrics for measurement of AI risks enumerated during the Map function are selected for implementation starting with the most significant AI risks. The risks or trustworthiness characteristics that will not – or cannot – be measured are properly documented.	• Measurements of identified risks are often more difficult for GPAIS than for smaller-scale or fixed-purpose AI systems, because of factors such as complexities, uncertainties, and emergent properties of GPAIS. However, it would not be appropriate to ignore identified risks just because measurement would be difficult, especially if the impacts could be severe or catastrophic.
	  ◦ For many factors it can be more appropriate to use qualitative assessment procedures, e.g., algorithmic impact assessments, human rights impact assessments, bug bounties, bias bounties, and red teams, because quantitative metrics for those factors might not be feasible or appropriate yet.
	  ◦ Plan to track and revisit identified risks, even if they cannot be measured quantitatively at this time, especially if the impacts could be severe or catastrophic. (See guidance in this document under Measure 3.2 on risk tracking approaches.)
		Section 3.2 of Barrett et al. (2022)
		
		Weidinger et al. (2023a,b)
		
		For AI red teaming general practices, including for LLMs, toxicity, and bias:
		• Casper et al. (2023a,b,c)
		• Google (2023b)
		• Ganguli, Lovitt et al (2022)
		• Su et al. (2023)
		
		For red teaming and dangerous capability evaluation of frontier models:
		• Ganguli, Lovitt et al. (2022)
		• OpenAI (2023a, pp. 15–16) and ARC Evals (2023a,b)
		• Anthropic (2023b,g)
		• Shevlane et al. (2023)
		• Kinniment et al. (2023)

Measure Category or Subcategory	Applicability and supplemental guidance for GPAIS	Resources
(continued)	• **Use red teams and adversarial testing as part of extensive interaction with GPAIS to identify dangerous capabilities, vulnerabilities, or other emergent properties of such systems.** Emergent properties are more likely with large-scale machine learning models than with smaller models, though it also might be more difficult or impossible to detect emergent dangerous capabilities or other characteristics of increasingly advanced AI (Hendrycks, Carlini et al. 2021 p. 7). Security vulnerabilities are typically inherent to currently available GPAIS, including in particular vulnerabilities to prompt injection attacks. (See, e.g., OWASP 2023a.) Red teaming can identify these weaknesses, though they are currently difficult to protect against. (See, e.g., Zou et al. 2023a,b.) ◦ For cutting-edge GPAIS, foundation models, and frontier models, characteristics that red teams should evaluate include: unacceptable-risk factors as outlined in guidance in this document under Map 1.5; and high-impact and catastrophic-harm factors as outlined in guidance in this document under Map 5.1, including dangerous capabilities such as advanced manipulation or deception. » The factors mentioned above include the following topics that are part of pre-release evaluation commitments by several frontier model developers (White House 2023a): – Dual-use potential for biological, chemical, and radiological risks – Cyber attack capabilities – Capacity to control physical systems – Capacity for self-replication – Societal risks, such as bias and discrimination ◦ For examples of procedures and lessons learned in red teaming of LLMs, see Ganguli, Lovitt et al. (2022) and Casper et al. (2023a,b,c). ◦ For examples of red team evaluations of dangerous capabilities in frontier models, see OpenAI (2023a, pp. 15–16), ARC Evals (2023a,b), Kinniment et al. (2023), and Anthropic (2023b,g); see also Shevlane et al. (2023) for related considerations. ◦ Consider automated generation of test cases as part of red team analyses. See, e.g., DeepMind's use of a language model for testing a version of the large language model Gopher (Perez, Huang et al. 2022) or Anthropic's model-written evaluations (Perez, Ringer et al. 2022a,b). ◦ **Partner with an independent red-teaming organization as appropriate.** OpenAI used the external red-teaming organization ARC Evals (which has expertise in safety of LLMs and other GPAIS) while developing GPT-4 and provided an overview of the emergent-properties testing ARC performed in the GPT-4 System Card (OpenAI 2023a, pp.15–16). AI companies have also participated in red-teaming events open to larger communities such as in DEF CON 31 which was open to attendees as well as civil society and community organizations (White House 2023b). » Several frontier-model developers have committed to external as well as internal red teaming (White House 2023a). ◦ Protect proprietary or unreleased foundation model weights as appropriate during red teaming to prevent unauthorized access or leaks of model weights. (For more on protecting proprietary or unreleased foundation model parameter weights, see guidance under Measure 2.7.) ◦ Grant red teams appropriate access to the final versions of foundation models before deployment. There might be cases where you grant a red team access to an early version of a model, and then perform additional fine-tuning on the model. In this case, go through the red teaming process again on the final version of the model to avoid missing dangerous emergent properties that might have been introduced during the fine-tuning process.	Language model benchmarks and other evaluations related to safety, ethics, and risks include: • BIG-bench "pro-social behavior" category of benchmark tasks (BIG-bench n.d.b, BIG-bench collaboration 2021, Srivastava et al. 2022) • Model-Written Evaluations "advanced-ai-risk," "sycophancy," and "winogender" datasets (Perez, Ringer et al. 2022a,b) For broader sets of language model evaluation and metrics, including of general knowledge and capabilities: • BIG-bench (BIG-bench collaboration 2021, Srivastava et al. 2022) • Evaluate library (Hugging Face 2022, Ngo, Thrush et al. 2022) in combination with datasets from BIG-bench or another dataset source • HELM (CRFM 2022, Liang et al. 2022) • LAMBADA (Paperno et al. 2016) • MMLU (Hendrycks, Burns et al. 2020a,b) • TriviaQA (Joshi et al. 2017a,b,c) • TruthfulQA (Lin et al. 2021a,b) • Model-Written Evaluations (Perez, Ringer et al. 2022a,b) For evaluation of computer programming (code generation) capabilities of language models: • APPS (Hendrycks, Basart et al. 2021a,b) • HumanEval (Chen et al. 2021)

Measure Category or Subcategory	Applicability and supplemental guidance for GPAIS	Resources
(continued)	◦ **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. ◦ When planning how much resources to devote to red teaming and adversarial testing, especially for frontier models, as points of comparison consider the levels of effort used in the examples cited in this section, e.g., the emergent-properties testing described in the GPT-4 System Card (OpenAI 2023a, pp.15–16). Following are additional guidelines: » "Following a well-defined research plan, subject matter and LLM experts will need to collectively spend substantial time (i.e. 100+ hours) working closely with models to probe for and understand their true capabilities in a target domain" (Anthropic 2023b). » "Auditors and red teamers need to be adequately resourced, informed, and granted sufficient time to conduct their work at a risk-appropriate level of rigor, not least due to the risk that shallow audits or red teaming efforts provide a sense of false assurance" (Anderljung, Barnhart et al. 2023 p. 26). • As part of critical thinking about benchmarks for GPAIS, consider that many such benchmarks are more focused on beneficial GPAIS capabilities and performance than on the risks when a GPAIS fails or is misused. However, capabilities evaluations can be an important part of assessing risks, e.g., for identifying dangerous capabilities that can be misused or abused. ◦ As part of criteria for use of benchmarks or other metrics for risk assessment purposes, and as part of communication of benchmarking results, clarify whether a specific benchmark directly measures a particular risk 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. • As part of language model trustworthiness and performance, which can include characteristics such as harmful bias and lack of robustness, consider using toolkits and benchmarks such as the following (with appropriate recognition of their limitations[26] in application contexts that might vary from the context of an AI system's training environment): ◦ BIG-bench (BIG-bench collaboration 2021, Srivastava et al. 2022) ◦ HELM (CRFM 2022, Liang et al. 2022) ◦ LAMBADA (Paperno et al. 2016) ◦ MMLU (Hendrycks, Burns et al. 2020a,b) ◦ See also resources under Measure 2 for specific trustworthiness characteristics, e.g., BBQ (Parrish et al. 2021a,b) as a resource for evaluating fairness and bias under Measure 2.11. • If specific benchmarks suggested in this section would have been appropriate but have become obsolete, then use analogous or related up-to-date benchmarks instead of or in addition to the older benchmarks. In the NIST AI RMF Playbook guidance for Measure 1.1, particularly valuable action and documentation items for GPAIS include: • *Establish approaches for detecting, tracking and measuring known risks, errors, incidents or negative impacts.*	For evaluation of mathematical capabilities of language models: • GSM8k (Cobbe et al. 2021a,b) • MATH (Hendrycks, Burns et al. 2021a,b) NIST (2023b)

26 On limitations of benchmarks, see e.g., Raji et al. (2021) and Schaeffer et al. (2023).

Measure Category or Subcategory	Applicability and supplemental guidance for GPAIS	Resources
(continued)	• *Identify transparency metrics to assess whether stakeholders have access to necessary information about system design, development, deployment, use, and evaluation.* • *Utilize accountability metrics to determine whether AI designers, developers, and deployers maintain clear and transparent lines of responsibility and are open to inquiries.* • *Document metric selection criteria and include considered but unused metrics.* • *Monitor AI system external inputs including training data, models developed for other contexts, system components reused from other contexts, and third-party tools and resources.* • *Report metrics to inform assessments of system generalizability and reliability.* • *Assess and document pre- vs post-deployment system performance. Include existing and emergent risks.* • *Document risks or trustworthiness characteristics identified in the Map function that will not be measured, including justification for non-measurement.* • *How will the appropriate performance metrics, such as accuracy, of the AI be monitored after the AI is deployed?* • *What testing, if any, has the entity conducted on the AI system to identify errors and limitations (i.e. manual vs automated, adversarial and stress testing)?*	
Measure 1.2: Appropriateness of AI metrics and effectiveness of existing controls are regularly assessed and updated, including reports of errors and potential impacts on affected communities.	In the NIST AI RMF Playbook guidance for Measure 1.2, particularly valuable action and documentation items for GPAIS include: • *Assess effectiveness of existing metrics and controls on a regular basis throughout the AI system lifecycle.* • *Document reports of errors, incidents and negative impacts and assess sufficiency and efficacy of existing metrics for repairs, and upgrades.* • *Develop new metrics when existing metrics are insufficient or ineffective for implementing repairs and upgrades.* • *Develop and utilize metrics to monitor, characterize and track external inputs, including any third-party tools.* • *Determine frequency and scope for sharing metrics and related information with stakeholders and impacted communities.* • *Utilize stakeholder feedback processes established in the Map function to capture, act upon and share feedback from end users and potentially impacted communities.* • *What metrics has the entity developed to measure performance of the AI system?* • *What is the justification for the metrics selected?*	NIST (2023b)

Measure Category or Subcategory	Applicability and supplemental guidance for GPAIS	Resources
Measure 1.3: Internal experts who did not serve as front-line developers for the system and/or independent assessors are involved in regular assessments and updates. Domain experts, users, AI actors external to the team that developed or deployed the AI system, and affected communities are consulted in support of assessments as necessary per organizational risk tolerance.	As part of assessments, make use of one or more red teams with expertise in safety of GPAIS as relevant. The teams should be separate enough from direct development operations of a GPAIS that they can provide relatively unbiased assessments of that GPAIS. (See also guidance in this document under Measure 1.1 for more detailed recommendations about using red teams and independent red-teaming organizations as independent assessors. See Govern 5.1 for more information about additional models of external feedback.) In the NIST AI RMF Playbook guidance for Measure 1.3, particularly valuable action and documentation items for GPAIS include: • *Evaluate TEVV processes regarding incentives to identify risks and impacts.* • *Utilize separate testing teams established in the Govern function (2.1 and 4.1) to enable independent decisions and course-correction for AI systems. Track processes and measure and document change in performance.* • *Assess independence and stature of TEVV and oversight AI actors, to ensure they have the required levels of independence and resources to perform assurance, compliance, and feedback tasks effectively.* • *Evaluate interdisciplinary and demographically diverse internal team established in Map 1.2.* • *Evaluate effectiveness of external stakeholder feedback mechanisms, specifically related to processes for eliciting, evaluating and integrating input from diverse groups.* • *What are the roles, responsibilities, and delegation of authorities of personnel involved in the design, development, deployment, assessment and monitoring of the AI system?* • *What type of information is accessible on the design, operations, and limitations of the AI system to external stakeholders, including end users, consumers, regulators, and individuals impacted by use of the AI system?*	NIST (2023b)
Measure 2: AI systems are evaluated for trustworthy characteristics.		
Measure 2.1: Test sets, metrics, and details about the tools used during TEVV are documented.	In the NIST AI RMF Playbook guidance for Measure 2.1, particularly valuable action and documentation items for GPAIS include: • *Leverage existing industry best practices for transparency and documentation of all possible aspects of measurements.* • *Regularly assess the effectiveness of tools used to document measurement approaches, test sets, metrics, processes and materials used.* • *Update the tools as needed.*	NIST (2023b)

Measure Category or Subcategory	Applicability and supplemental guidance for GPAIS	Resources
Measure 2.2: Evaluations involving human subjects meet applicable requirements (including human subject protection) and are representative of the relevant population.	In the NIST AI RMF Playbook guidance for Measure 2.2, particularly valuable action and documentation items for GPAIS include: • *Follow human subjects research requirements as established by organizational and disciplinary requirements, including informed consent and compensation, during dataset collection activities.* • *Follow intellectual property and privacy rights related to datasets and their use, including for the subjects represented in the data.* • *Use informed consent for individuals providing data used in system testing and evaluation.* • *How has the entity identified and mitigated potential impacts of bias in the data, including inequitable or discriminatory outcomes?* • *To what extent are the established procedures effective in mitigating bias, inequity, and other concerns resulting from the system?* • *If human subjects were used in the development or testing of the AI system, what protections were put in place to promote their safety and wellbeing?*	NIST (2023b)
Measure 2.3: AI system performance or assurance criteria are measured qualitatively or quantitatively and demonstrated for conditions similar to deployment setting(s). Measures are documented.	In the NIST AI RMF Playbook guidance for Measure 2.3, particularly valuable action and documentation items for GPAIS include: • *Conduct regular and sustained engagement with potentially impacted communities* • *Maintain a demographically diverse and multidisciplinary and collaborative internal team* • *Evaluate feedback from stakeholder engagement activities, in collaboration with human factors and socio-technical experts.* • *Measure AI systems prior to deployment in conditions similar to expected scenarios.* • *What testing, if any, has the entity conducted on the AI system to identify errors and limitations (i.e. adversarial or stress testing)?* (See also guidance in this document for Govern 2.1 regarding roles for upstream developers as well as downstream developers and deployers, and see guidance in this document under Measure 1.1 on approaches to measuring identified risks for GPAIS.)	NIST (2023b)
Measure 2.4: The functionality and behavior of the AI system and its components – as identified in the Map function – are monitored when in production.	In the NIST AI RMF Playbook guidance for Measure 2.4, particularly valuable action and documentation items for GPAIS include: • *Monitor for anomalies using approaches such as control limits, confidence intervals, integrity constraints and ML algorithms. When anomalies are observed, consider error propagation and feedback loop risks.* • *Collect uses cases from the operational environment for system testing and monitoring activities in accordance with organizational policies and regulatory or disciplinary requirements (e.g. informed consent, institutional review board approval, human research protections).* • *How will the appropriate performance metrics, such as accuracy of the AI, be monitored after the AI is deployed?* (See guidance in this document under Govern 2.1 regarding roles for upstream developers as well as downstream developers and deployers, and see guidance in this document under Measure 1.1 on approaches to measuring identified risks for GPAIS.)	NIST (2023b)

Measure Category or Subcategory	Applicability and supplemental guidance for GPAIS	Resources

Measure 2.5: The AI system to be deployed is demonstrated to be valid and reliable. Limitations of the generalizability beyond the conditions under which the technology was developed are documented.	In the NIST AI RMF Playbook guidance for Measure 2.5, particularly valuable action and documentation items for GPAIS include:
	• *Establish or identify, and document approaches to measure forms of validity, including:*
	  ◦ *construct validity (the test is measuring the concept it claims to measure)*
	  ◦ *internal validity (relationship being tested is not influenced by other factors or variables)*
	  ◦ *external validity (results are generalizable beyond the training condition)*
	  ◦ *the use of experimental design principles and statistical analyses and modeling.*
	• *Establish or identify, and document robustness measures.*
	• *Establish or identify, and document reliability measures.*
	• *Establish practices to specify and document the assumptions underlying measurement models to ensure proxies accurately reflect the concept being measured.*
	• *What testing, if any, has the entity conducted on the AI system to identify errors and limitations (i.e. adversarial or stress testing)?*
	• *To what extent are the established procedures effective in mitigating bias, inequity, and other concerns resulting from the system?*
	
	(See also guidance in this document for Govern 2.1 regarding roles for upstream developers as well as downstream developers and deployers, guidance in this document under Measure 1.1 on approaches to measuring identified risks for GPAIS, and guidance in this document under Map 1.3 and Map 5.1 for qualitative approaches to characterizing AI system objectives mis-specification or goal misgeneralization.)
		For LLMs:
		• TruthfulQA (Lin et al. 2021a,b)
		• LAMBADA (Paperno et al. 2016)
		• MMLU (Hendrycks, Burns et al. 2020)
		• Winogender (Rudinger et al. 2019)
		• BIG-bench "pro-social behavior" category of benchmark tasks (BIG-bench n.d.b, BIG-bench collaboration 2021, Srivastava et al. 2022)
		• Model-Written Evaluations "advanced-ai-risk," "sycophancy" and "winogender" datasets (Perez, Ringer et al. 2022a,b)
		
		NIST (2023b)

Measure 2.6: The AI system is evaluated regularly for safety risks – as identified in the Map function. The AI system to be deployed is demonstrated to be safe, its residual negative risk does not exceed the risk tolerance, and it can fail safely, particularly if made to operate beyond its knowledge limits. Safety metrics reflect system reliability and robustness, real-time monitoring, and response times for AI system failures.	As part of safety evaluations of GPAIS:
	• Perform red teaming and adversarial testing of safety aspects of GPAIS; for frontier models this testing should include dangerous-capability evaluations. (See also guidance in this document under Measure 1.1 on red teaming and dangerous capability evaluations.)
	
	In the NIST AI RMF Playbook guidance for Measure 2.6, particularly valuable action and documentation items for GPAIS include:
	• *Thoroughly measure system performance in development and deployment contexts, and under stress conditions.*
	  ◦ *Employ test data assessments and simulations before proceeding to production testing. Track multiple performance quality and error metrics.*
	  ◦ *Stress-test system performance under likely scenarios (e.g., concept drift, high load) and beyond known limitations, in consultation with domain experts.*
	  ◦ *Test the system under conditions similar to those related to past known incidents or near-misses and measure system performance and safety characteristics.*
	• *Measure and monitor system performance in real-time to enable rapid response when AI system incidents are detected.*
	• *Document, practice and measure incident response plans for AI system incidents, including measuring response and down times.*
	• *What testing, if any, has the entity conducted on the AI system to identify errors and limitations (i.e. adversarial or stress testing)?*
		For red teaming and dangerous capability evaluation of frontier models:
		• OpenAI (2023a, pp. 15–16) and ARC Evals (2023a,b)
		• Kinniment et al. (2023)
		• Shevlane et al. (2023)
		
		For red teaming LLMs and toxicity:
		• Casper et al. (2023a,b,c)
		
		For LLM truthfulness and toxicity:
		• ToxiGen (Hartvigsen et al. 2022)
		• TruthfulQA (Lin et al. 2021a,b)
		
		AIID (n.d.)
		
		NIST (2023b)

Measure Category or Subcategory	Applicability and supplemental guidance for GPAIS	Resources
(continued)	• *Did you establish mechanisms that facilitate the AI system's auditability (e.g. traceability of the development process, the sourcing of training data and the logging of the AI system's processes, outcomes, positive and negative impact)?* ◦ *For some GPAIS (e.g., using models run on central servers accessed through APIs), these can include data mining of usage metrics, audit logs, etc. as appropriate to identify anomalous conditions that users encounter but might not report.* • *Did you ensure that the AI system can be audited by independent third parties?* • *Did you establish a process for third parties (e.g. suppliers, end-users, subjects, distributors/vendors or workers) to report potential vulnerabilities, risks or biases in the AI system?*	
Measure 2.7: AI system security and resilience – as identified in the Map function – are evaluated and documented.	Use information security measures to assess and assure model weight security (specifically, integrity and confidentiality) as part of preventing misuse or abuse of models. This is particularly valuable for frontier models, for which public release of model weights could enable misuse with particularly high-consequence impacts. • Anthropic has announced their frontier-model security practices include requirements for multi-party authorization for access to frontier model development and deployment systems, and secure development and supply chain practices, including chain of custody (Anthropic 2023a). As a general guideline for information system **security expectations for protecting the integrity and confidentiality of proprietary or unreleased foundation model parameter weights**, foundation model developers should implement the NIST Cybersecurity Framework (NIST 2018), or an approximate equivalent such as NIST SP 800-171 or ISO/IEC 27001, with at least the following security controls or approximate equivalents:[27] • 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.[28] • 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. As part of security evaluations of GPAIS: • Perform red teaming and adversarial testing of security aspects of GPAIS. (See also guidance in this document under Measure 1.1 on red teaming and adversarial testing.)	NIST (2023b) On baseline expectations for information system security for foundation model developers: • NIST Cybersecurity Framework (NIST 2018) • NIST SP 800-53B (NIST 2020a) • NIST SP 800-171 (NIST 2020b) • NIST SP 800-172 (NIST 2021) • ISO/IEC (2022) • Anthropic (2023a,g) On security vulnerabilities and mitigations for LLMs and other types of ML models: • ENISA (2021, 2023) • Oprea and Vassilev (2023) • OWASP (2023a,b) • Barrett, Boyd et al. (2023) • ATLAS (MITRE n.d.) • TrojAI (Karra et al. 2020, NIST n.d.b)

27 For approximate equivalents, see, e.g., the NIST (2020b) mappings of controls between NIST SP 800-171 and NIST 800-53 and ISO/IEC 27001; the NIST (2021) mapping of controls between NIST SP 800-172 and NIST SP 800-53; the NIST (2020c) mappings of controls between the NIST Cybersecurity Framework and NIST SP 800-53; the NIST (2023e) mapping of controls between NIST SP 800-53 and ISO/IEC 27001, and the CIS (n.d.) mapping of controls between NIST SP 800-53 and CIS Critical Security Controls.

28 Several frontier-model developers have committed to investing in cybersecurity and insider-threat controls for a high level of protection of proprietary and unreleased frontier-model weights. "This includes limiting access to model weights to those whose job function requires it and establishing a robust insider threat detection program consistent with protections provided for their most valuable intellectual property and trade secrets. In addition, it requires storing and working with the weights in an appropriately secure environment to reduce the risk of unsanctioned release" (White House 2023a, p.3).

Measure Category or Subcategory	Applicability and supplemental guidance for GPAIS	Resources
(continued)	• 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. Tools to consider using include TrojAI (Karra et al. 2020, NIST n.d.b); see also Oprea and Vassilev (2023). • Engage in continuous monitoring, vulnerability disclosure, and bug bounty programs for GPAIS to identify novel security vulnerabilities. • Track uncovered security vulnerabilities in other GPAIS, including open-source foundation models, which may be transferable to other models. (See, e.g., Zou et al. 2023a,b.) In the NIST AI RMF Playbook guidance for Measure 2.7, particularly valuable action and documentation items for GPAIS include: • *Establish and track AI system security tests and metrics (e.g., red-teaming activities, frequency and rate of anomalous events, system down-time, incident response times, time-to-bypass, etc.).* • *Use red-team exercises to actively test the system under adversarial or stress conditions, measure system response, assess failure modes or determine if system can return to normal function after an unexpected adverse event.* • *Document red-team exercise results as part of continuous improvement efforts, including the range of security test conditions and results.* • *Verify that information about errors and attack patterns is shared with incident databases, other organizations with similar systems, and system users and stakeholders (see also related guidance under Manage 4.1).* • *Develop and maintain information sharing practices with AI actors from other organizations to learn from common attacks.* • *Verify that third party AI resources and personnel undergo security audits and screenings. Risk indicators may include failure of third parties to provide relevant security information.* • *Utilize watermarking technologies as a deterrent to data and model extraction attacks.*	For a range of LLM red-teaming approaches with security implications: • Ganguli, Lovitt et al. (2022) • Casper et al. (2023a,b,c) • Zou et al. (2023a,b) • OpenAI (2023a, pp. 15-16) and ARC Evals (2023a,b) • Kinniment et al. (2023) • Anthropic (2023b,g) • Shevlane et al. (2023)

Measure 2.8:
Risks associated with transparency and accountability – as identified in the Map function – are examined and documented.
	In the NIST AI RMF Playbook guidance for Measure 2.8, particularly valuable action and documentation items for GPAIS include:
	• *Instrument the system for measurement and tracking, e.g., by maintaining histories, audit logs, and other information that can be used by AI actors to review and evaluate possible sources of error, bias, or vulnerability.*
	• *Track, document, and measure organizational accountability regarding AI systems via policy exceptions and escalations, and document "go" and "no/go" decisions made by accountable parties.*
	• *Track and audit the effectiveness of organizational mechanisms related to AI risk management, including:*
	  ◦ *Lines of communication between AI actors, executive leadership, users, and impacted communities.*
	  ◦ *Roles and responsibilities for AI actors and executive leadership.*
	  ◦ *Organizational accountability roles, e.g., chief model risk officers, AI oversight committees, responsible or ethical AI directors, etc.*
	
	Document organizational transparency and disclosure mechanisms to inform users or allow users to check whether they are interacting with, or observing content created by, a generative AI system. See, e.g., Partnership on AI's Responsible Practices for Synthetic Media (PAI 2023a), as well as CAI (2023) and C2PA (2023).
	
	(See also guidance in this document under Govern 2.1 on roles for GPAIS upstream and downstream developers, and under Manage 1.3 on transparency and disclosure.)
		PAI (2023a)
		CAI (2023)
		C2PA (2023)
		Solaiman (2023)
		NIST (2023b)

Measure Category or Subcategory	Applicability and supplemental guidance for GPAIS	Resources
Measure 2.9: The AI model is explained, validated, and documented, and AI system output is interpreted within its context – as identified in the Map function – to inform responsible use and governance.	It is critical to ensure that users know how to interpret system behavior and outputs, including the limitations of both the system and any explanations provided. However, explainability and interpretability are often extremely limited for LLMs and other GPAIS with deep-learning architectures. These systems can be inappropriate for applications requiring better explainability and interpretability. For some increasingly capable GPAIS, the reliability of some techniques (such as RLHF) for aligning GPAIS behavior with human values or intentions could depend on being combined with sufficient interpretability methods to prevent "deceptive alignment" (Hubinger et al. 2019, Ngo, Chan et al. 2022). • While interpretability techniques are not yet sufficient to assess risks such as hidden failures of RLHF for GPAIS alignment, developers of GPAIS (especially frontier models) should include such risks in a risk register or other tool for tracking identified risks that are difficult to assess. (See related guidance in this document under Measure 3.2.) In the NIST AI RMF Playbook guidance for Measure 2.9, particularly valuable action and documentation items for GPAIS include: • *What type of information is accessible on the design, operations, and limitations of the AI system to external stakeholders, including end users, consumers, regulators, and individuals impacted by use of the AI system?* (See also guidance in this document for Govern 2.1 regarding roles for upstream developers as well as downstream developers and deployers, and see guidance in this document under Measure 1.1 on approaches to measuring identified risks for GPAIS.)	Mitchell et al. (2019) NIST (2023b)
Measure 2.10: Privacy risk of the AI system – as identified in the Map function – is examined and documented.	Privacy challenges for GPAIS 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. In the NIST AI RMF Playbook guidance for Measure 2.10, particularly valuable action and documentation items for GPAIS include: • *Document collection, use, management, and disclosure of personally sensitive information in datasets, in accordance with privacy and data governance policies.* • *Establish and document protocols (authorization, duration, type) and access controls for training sets or production data containing personally sensitive information, in accordance with privacy and data governance policies.* • *Monitor internal queries to production data for detecting patterns that isolate personal records.* • *Did your organization implement accountability-based practices in data management and protection (e.g. the PDPA and OECD Privacy Principles)?* • *What assessments has the entity conducted on data security and privacy impacts associated with the AI system?* Additional valuable steps to consider include: • Enable people to consent to the uses of their data and opt out of the uses of their data. • Notify users and impacted communities about privacy or security breaches. (See also guidance in this document for Govern 2.1 regarding roles for upstream developers as well as downstream developers and deployers, and see guidance in this document under Measure 1.1 on approaches to measuring identified risks for GPAIS.)	NIST (2023b)

Measure Category or Subcategory	Applicability and supplemental guidance for GPAIS	Resources

Measure 2.11: Fairness and bias – as identified in the Map function – are evaluated and results are documented.	In the NIST AI RMF Playbook guidance for Measure 2.11, particularly valuable action and documentation items for GPAIS include:
	• *Understand and consider sources of bias in training and TEVV data:*
	  ◦ *Differences in distributions of outcomes across and within groups, including intersecting groups.*
	  ◦ *Completeness, representativeness and balance of data sources.*
	  ◦ *Identify input data features that may serve as proxies for demographic group membership (i.e., credit score, ZIP code) or otherwise give rise to emergent bias within AI systems.*
	  ◦ *Forms of systemic bias in images, text (or word embeddings), audio or other complex or unstructured data.*
	• *Leverage impact assessments to identify and classify system impacts and harms to end users, other individuals, and groups with input from potentially impacted communities.*
	• *Identify the classes of individuals, groups, or environmental ecosystems which might be impacted through direct engagement with potentially impacted communities.*
	• *Collect and share information about differences in outcomes for the identified groups.*
	• *How has the entity identified and mitigated potential impacts of bias in the data, including inequitable or discriminatory outcomes?*
	
	Additional valuable steps include:
	• Review AI system development and uses for potential threats to human rights, dignity, or wellbeing.
	• Ensure the AI system's user interface is usable by those with special needs or disabilities, or those at risk of exclusion.
	• Determine methods to distribute the benefits of the system widely and equitably.
	
	(See also guidance in this document under Map 5.1 on identifying potential large-scale harms from correlated bias across large numbers of people or a large fraction of a group or a society's population.)
		For LLMs:
		• BBQ (Parrish et al. 2021a,b)
		• Winogender Schemas (Rudinger et al, 2019)
		• ToxiGen (Hartvigsen et al. 2022)
		• TruthfulQA (Lin et al. 2021a,b)
		• BOLD (Dhamala 2021)
		• Su et al. (2023)
		
		Aequitas (Saleiro 2019)
		AIFairness 360 (Bellamy 2018)
		Fairlearn (Fairlearn Contributors 2023)
		
		NIST (2023b)
		Schwartz et al. (2022)

Measure 2.12: Environmental impact and sustainability of AI model training and management activities – as identified in the Map function – are assessed and documented.	Environmental impact assessment by GPAIS developers should include estimating the environmental impact of large-scale ML model training. • Relevant tools, resources, and examples include ML CO_2 Impact (Schmidt et al. 2019), Lacoste et al. (2019), OECD (2022b), and Luccioni et al. (2022). • Assessment of environmental impacts is particularly important for LLMs and other large-scale ML-based AI systems, which typically have much larger model-training environmental impacts than smaller-scale ML models (Bender et al. 2021).	Schmidt et al. (2019) Lacoste et al. (2019) OECD (2022b) Luccioni et al. (2022) NIST (2023b)

Measure Category or Subcategory	Applicability and supplemental guidance for GPAIS	Resources
Measure 2.13: Effectiveness of the employed TEVV metrics and processes in the Measure function are evaluated and documented.	In the NIST AI RMF Playbook guidance for Measure 2.13, particularly valuable action and documentation items for GPAIS include: • *Assess effectiveness of metrics for identifying and measuring risks.*	NIST (2023b)
Measure 3: Mechanisms for tracking identified AI risks over time are in place.		

Measure 3.1: Approaches, personnel, and documentation are in place to regularly identify and track existing, unanticipated, and emergent AI risks based on factors such as intended and actual performance in deployed contexts.	In the NIST AI RMF Playbook guidance for Measure 3.1, particularly valuable action and documentation items for GPAIS include:
	• *Assess effectiveness of metrics for identifying and measuring emergent risks.*
	• *To what extent can users or parties affected by the outputs of the AI system test the AI system and provide feedback?*
	
	Additional valuable steps to consider include:
	• Consider steps to identify or assess longer-term impacts or use longer time horizons (longer than would be typical for smaller-scale, fixed-purpose AI systems), and 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).
	  ◦ If your AI system is deployed for a long period of time, then:
	  » What additional impacts would you expect?
	  » Which impacts would you expect to have greater magnitude?
	• 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. (These can be particularly important for foundation models, which often have emergent capabilities and other emergent properties that are not identified in earlier-stage testing.) To identify trigger events, consider questions such as:
	  ◦ What if monitoring indicates one of your risk-mitigation controls is not working as expected? (Consider this, as applicable, for each relevant risk-mitigation control.)
	  ◦ What if AI capability developments occur that are not expected until further into the future, such as availability of much more powerful AI systems or computing resources to train and run AI systems, or demonstration of new emergent capabilities (e.g., via new prompts) that were not identified in earlier-stage testing?
	  ◦ What if a near-miss incident occurs in a critical system or process? Does your organization have procedures for near-miss incident identification, analysis, tracking, and information sharing? Does your organization also monitor the AIID or other sources for near-miss incident reports on other organizations' systems?
		AIID (n.d.)
		Section 3.2 of Barrett et al. (2022)
		NIST (2023b)

Measure Category or Subcategory	Applicability and supplemental guidance for GPAIS	Resources

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.	**Use appropriate mechanisms for tracking identified risks, even if only characterizing them qualitatively and even if the risks are difficult to assess.** This is particularly important for foundation models, because of their potential scale of impact, and their potential for emergent properties or other novel risks.
	• Consider tracking identified risks (including difficult-to-assess risks) using a risk register. (For more on risk registers, see, e.g., ISO Guide 73 Section 3.8.2.4, PMI 2017 p. 417, and Stine et al. 2020.)
	• When developing frontier models with unprecedented capabilities, failure modes, and other emergent properties, it is especially valuable to use red teams and adversarial testing prior to deployment. See related guidance in this document under Measure 1.1.
	• **Risk tracking should include ongoing monitoring of newly identified capabilities and limitations of deployed GPAIS.** These efforts can include monitoring use of the models through APIs, and monitoring publications or online forums that discuss new uses of the models. "If significant information on model capabilities is discovered post-deployment, risk assessments should be repeated, and deployment safeguards updated" (Anderljung, Barnhart et al. 2023).
	
	In the NIST AI RMF Playbook guidance for Measure 3.2, particularly valuable action and documentation items for GPAIS include:
	• *Establish processes for tracking emergent risks that may not be measurable with current approaches. Some processes may include:*
	  ◦ *Recourse mechanisms for faulty AI system outputs.*
	  ◦ *Bug bounties.*
	  ◦ *Human-centered design approaches.*
	  ◦ *User-interaction and experience research.*
	  ◦ *Participatory stakeholder engagement with affected or potentially impacted individuals and communities.*
	• *Determine and document the rate of occurrence and severity level for complex or difficult-to-measure risks when:*
	  ◦ *Prioritizing new measurement approaches for deployment tasks.*
	  ◦ *Allocating AI system risk management resources.*
	  ◦ *Evaluating AI system improvements.*
	  ◦ *Making go/no-go decisions for subsequent system iterations.*
		Section 3.2 of Barrett et al. (2022)
		
		NIST (2023b)
		
		On bug bounties and bias bounties:
		• Globus-Harris et al. (2022)
		• Kenway et al. (2022)
		• OpenAI (2023c)

Measure 3.3: Feedback processes for end users and impacted communities to report problems and appeal system outcomes are established and integrated into AI system evaluation metrics.	In the NIST AI RMF Playbook guidance for Measure 3.3, particularly valuable action and documentation items for GPAIS include: • *To what extent can users or parties affected by the outputs of the AI system test the AI system and provide feedback?* • *How easily accessible and current is the information available to external stakeholders?* • *What type of information is accessible on the design, operations, and limitations of the AI system to external stakeholders, including end users, consumers, regulators, and individuals impacted by use of the AI system?*	NIST (2023b)

Measure Category or Subcategory	Applicability and supplemental guidance for GPAIS	Resources
Measure 4: Feedback about efficacy of measurement is gathered and assessed.		
Measure 4.1: Measurement approaches for identifying AI risks are connected to deployment context(s) and informed through consultation with domain experts and other end users. Approaches are documented.	For GPAIS developers, GPAIS "users" include downstream developers as well as the end users of applications built on GPAIS platforms. Downstream developers typically have the most direct interactions with end users in particular deployment contexts. However, it can be valuable for upstream GPAIS developers to provide mechanisms for feedback from end users or other AI actors, as well as from downstream developers. (See also guidance in this document under Govern 2.1, regarding roles for GPAIS developers, e.g., on performing testing during GPAIS development or other testing that requires direct access to the system, as well as downstream developers and deployers, e.g., on performing testing of end-use applications built on a GPAIS and testing appropriate for that application context.)	NIST (2023b)
Measure 4.2: Measurement results regarding AI system trustworthiness in deployment context(s) and across the AI lifecycle are informed by input from domain experts and relevant AI actors to validate whether the system is performing consistently as intended. Results are documented.	In the NIST AI RMF Playbook guidance for Measure 4.2, particularly valuable action and documentation items for GPAIS include: • *Integrate feedback from end users, operators, and affected individuals and communities from Map function as inputs to assess AI system trustworthiness characteristics. Ensure both positive and negative feedback is being assessed.* • *Evaluate feedback in connection with AI system trustworthiness characteristics from Measure 2.5 to 2.11.* • *Consult AI actors in impact assessment, human factors and socio-technical tasks to assist with analysis and interpretation of results.* When considering what types of domain experts to use in reviewing information on identified risks, consider including personnel recommended for risk identification per guidance in this document under Govern 3.1, such as social scientists for perspective on structural or systemic risks. (See also guidance in this document under Govern 2.1, regarding roles for GPAIS developers, e.g., on performing testing during GPAIS development or other testing that requires direct access to the system, as well as downstream developers and deployers, e.g., on performing testing of end-use applications built on a GPAIS and testing appropriate for that application context.)	NIST (2023b)
Measure 4.3: Measurable performance improvements or declines based on consultations with relevant AI actors, including affected communities, and field data about context-relevant risks and trustworthiness characteristics are identified and documented.	In the NIST AI RMF Playbook guidance for Measure 4.3, particularly valuable action and documentation items for GPAIS include: • *Develop baseline quantitative measures for trustworthy characteristics.* • *Delimit and characterize baseline operation values and states.* • *Utilize qualitative approaches to augment and complement quantitative baseline measures, in close coordination with impact assessment, human factors and socio-technical AI actors.* • *Monitor and assess measurements as part of continual improvement to identify potential system adjustments or modifications.* (See also guidance in this document under Govern 2.1, regarding roles for GPAIS developers, e.g., on performing testing during GPAIS development or other testing that requires direct access to the system, as well as downstream developers and deployers, e.g., on performing testing of end-use applications built on a GPAIS and testing appropriate for that application context.)	NIST (2023b)

3.4 GUIDANCE FOR NIST AI RMF MANAGE SUBCATEGORIES

Table 4: Guidance for NIST AI RMF Manage Subcategories

Manage Category or Subcategory	Applicability and supplemental guidance for GPAIS	Resources
Manage 1: AI risks based on assessments and other analytical output from the Map and Measure functions are prioritized, responded to, and managed.		

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.	As part of considerations of the intended purpose of a GPAIS (if any), in addition to any originally intended use cases, include consideration of other identified potential use cases; see related guidance in this document under Map 1.1. This is particularly important for GPAIS, which can have large numbers of uses.
	
	When making go/no-go decisions, especially on whether to proceed on major stages or investments for development or deployment of cutting-edge large-scale GPAIS:
	• 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.**
	• See guidance in this document under Map 1.5 on organizational risk tolerances, especially: **Set policies on unacceptable-risk thresholds for GPAIS development and GPAIS deployment to include prevention of risks with substantial probability of inadequately-mitigated catastrophic outcomes.**
	  ◦ As previously mentioned, **the NIST AI RMF 1.0 strongly suggests considering catastrophic risks as unacceptable**: "In cases where an AI system presents unacceptable negative risk levels – such as **where significant negative impacts are imminent, severe harms are actually occurring, or catastrophic risks are present – development and deployment should cease in a safe manner until risks can be sufficiently managed** [emphasis added]" (NIST 2023a, p.8).
	• Check or update, and incorporate, guidance in this document under Map 1.5, especially: **Identify whether a GPAIS could lead to catastrophic impacts**.
	
	In the NIST AI RMF Playbook guidance for Manage 1.1, particularly valuable action and documentation items for GPAIS include:
	  ◦ *Utilize TEVV outputs from map and measure functions when considering risk treatment.*
	  ◦ *Regularly track and monitor negative risks and benefits throughout the AI system lifecycle including in post-deployment monitoring.*
		NIST (2023b)

Manage Category or Subcategory	Applicability and supplemental guidance for GPAIS	Resources
Manage 1.2: Treatment of documented AI risks is prioritized based on impact, likelihood, and available resources or methods.	When prioritizing identified GPAIS risks: • Incorporate both impact and likelihood estimates as appropriate. See guidance in this document under Map 5.1 on assessing the magnitude of potential impacts of GPAIS risks. • Consider (i.e., do not ignore) risks that are difficult to assess, such as potential for emergent properties of GPAIS. See guidance in this document under Measure 3.2 on tracking risks that are difficult to assess. When considering available resources for risk treatment, see guidance in this document under Govern 2.1, e.g., for other risk assessment and risk management tasks for which upstream developers have substantially greater information and capability than others in the value chain, such as for assessing and mitigating early-stage GPAIS development risks. In the NIST AI RMF Playbook guidance for Manage 1.2, particularly valuable action and documentation items for GPAIS include: • *Regularly review risk tolerances and re-calibrate, as needed, in accordance with information from AI system monitoring and assessment.* (See also guidance on setting risk tolerances, in this document under Map 1.5.)	NIST (2023b)
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.	After identifying and analyzing use cases and misuse cases of an AI system (per "Map" function guidance): • For each identified potential use or misuse (or category of use or misuse) of an AI system: ◦ **Define and communicate to key stakeholders whether any potential use cases (or categories of use cases) would be disallowed/unacceptable, or would be treated as "high risk"** or another category for which your organization would provide specific risk management guidance or other risk mitigation measures. » E.g., it can be appropriate to consider whether any potential uses would be regarded under the EU AI Act as falling into one of the following risk categories: "unacceptable risk," "high risk," or "low or minimal risk," per draft EU AI Act language (EU 2021a Section 5.2.2). For example, AI systems would fall in the "unacceptable risk" category if their use would violate fundamental rights. » OpenAI recommends publishing usage guidelines and terms of use as part of prevention of misuse of LLMs (Cohere, OpenAI and AI21 Labs 2022). OpenAI's 2019 announcement of GPT-2 included listing several categories of potential misuse cases (OpenAI 2019a), which apparently informed their decisions on disallowed/unacceptable use-case categories of applications based on GPT-3 (OpenAI 2020). » Options for communicating whether uses would be disallowed or out of scope can include model cards (Mitchell et al. 2019) or related frameworks, as well as Responsible AI Licenses, or RAIL. Hugging Face and BigScience's release of the BLOOM LLM included a RAIL with usage restrictions disallowing various types of misuse (RAIL n.d., Contractor et al. 2022). Google lists categories of prohibited uses for its generative AI services (Google 2023a). **Regarding pre-design and planning:** • 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.	Barrett et al. (2022) Mitchell et al. (2019) Moës et al. (2023) Schuett et al. (2023) Solaiman (2023) PAI (2023a) NIST (2023b)

Manage Category or Subcategory	Applicability and supplemental guidance for GPAIS	Resources
(continued)	◦ While data poisoning can be an issue for any machine learning model, this might be particularly challenging for training cutting-edge large models; often training of the largest new models has relied heavily on large-scale, uncurated internet-scrape datasets (Bommasani et al. 2021 p. 106). ◦ As part of data curation, ensure that any data with the BIG-bench canary GUID is excluded from training data. (See, e.g., documentation at BIG-bench n.d.) **Regarding design and development:** • **See guidance in this document under Measure 2.7** on guidelines **for protecting the integrity and confidentiality of proprietary or unreleased foundation model parameter weights**. • Consider disallowing open-ended learning with live web access; instead consider measures such as disallowing access to web forms (Nakano et al. 2021), disallowing HTTP POST requests, etc. • **Increase the amount of compute (computing power) spent training frontier models only incrementally (e.g., by not more than three times between each increment) as part of identification and management of risks of emergent properties.** ◦ Often it is difficult to predict what failure modes machine learning models will have, what their performance will be, or what capabilities they will have. Machine learning systems are self-organizing systems that learn many patterns or features without explicit instruction. Incremental scaling-up approaches provide more opportunities for red-team monitors to identify emergent properties at an early or partially emergent stage, when responses to identified emergent properties might be more feasible and effective. (For related discussion of emergent properties see, e.g., Section 3 of Hendrycks, Carlini et al. 2021, and Bommasani et al. 2021.) Incremental scaling can also be a valuable part of predicting large-scale model performance, as with GPT-4 (OpenAI 2023b). • **Test frontier models after each incremental increase of compute, data, or model size for model training.** If a large incremental increase (e.g., three times or more compute, or two times or more data or model parameters)[29] was used in a particular model training increment compared to the previous model training increment, it will be particularly important for the new model to be heavily probed/monitored/stress-tested using detailed analysis processes (including red team methods) to identify emergent properties such as capabilities and failure modes. ◦ Anthropic's Responsible Scaling Policy includes model evaluations at "every 4x increase in effective compute" during training of models approximately equivalent to the mid-2023 state of the art (Anthropic 2023g, p. 11). **Regarding test and evaluation:** • See guidance in this document under Measure, including under Measure 1.1 on red teaming. • After training and before deployment, probe/monitor/stress cutting-edge GPAIS using detailed analysis processes (including or extending standard cybersecurity red team methods) to achieve testing objectives including: ◦ Testing for unintended toxic and harmful content and/or dangerous errors (e.g., inaccurate medical information). ◦ Identifying emergent properties such as new capabilities and failure modes.	

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

Manage Category or Subcategory	Applicability and supplemental guidance for GPAIS	Resources
(continued)	**To further improve reliability in design and development, test and evaluation, and deployment:** • Consider approaches to design, testing, and deployment so that AI systems possess the minimum necessary capabilities for high-reliability operation and not more 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. **On transparency and disclosure of generative AI outputs:** • Implement transparency and disclosure mechanisms to inform users or allow users to check whether they are interacting with, or observing content created by, a generative AI system. See, e.g., Partnership on AI's Responsible Practices for Synthetic Media (PAI 2023a). Additional valuable steps include: • Determine a strategy to safely and appropriately release the AI system, and determine what protections might be necessary to prevent harm or misuse. (See, e.g., Solaiman 2023; see also guidance in this document under Manage 2.4, including on open-source and open-access release.) • Allow people to opt out of the use of the AI system. • Support independent third-party auditing and evaluation of the AI system. • Provide redress to people who are negatively affected by the use of the AI system. In the NIST AI RMF Playbook guidance for Manage 1.3, particularly valuable action and documentation items for GPAIS include: • *Document procedures for acting on AI system risks related to trustworthiness characteristics.* • *Prioritize risks involving physical safety, legal liabilities, regulatory compliance, and negative impacts on individuals, groups, or society.* • *Identify risk response plans and resources and organizational teams for carrying out response functions.*	
Manage 1.4: Negative residual risks (defined as the sum of all unmitigated risks) to both downstream acquirers of AI systems and end users are documented.	In the NIST AI RMF Playbook guidance for Manage 1.4, particularly valuable action and documentation items for GPAIS include: • *Document residual risks within risk response plans, denoting risks that have been accepted, transferred, or subject to minimal mitigation.* • *Establish procedures for disclosing residual risks to relevant downstream AI actors.* • *Inform relevant downstream AI actors of requirements for safe operation, known limitations, and suggested warning labels as identified in MAP 3.4.* (See also guidance in this document under Govern 2.1 and Govern 4.2 on documenting and communicating risks to downstream actors and other relevant stakeholders as appropriate.)	NIST (2023b)

Manage Category or Subcategory	Applicability and supplemental guidance for GPAIS	Resources
Manage 2: Strategies to maximize AI benefits and minimize negative impacts are planned, prepared, implemented, documented, and informed by input from relevant AI actors.		

Manage 2.1: Resources required to manage AI risks are taken into account – along with viable non-AI alternative systems, approaches, or methods – to reduce the magnitude or likelihood of potential impacts.	In the NIST AI RMF Playbook guidance for Manage 2.1, particularly valuable action and documentation items for GPAIS include:
	• *Plan and implement risk management practices in accordance with established organizational risk tolerances.*
	• *Verify risk management teams are resourced to carry out functions, including:*
	  ◦ *Establishing processes for considering methods that are not automated; semi-automated; or other procedural alternatives for AI functions.*
	  ◦ *Enhance AI system transparency mechanisms for AI teams.*
	  ◦ *Enable exploration of AI system limitations by AI teams.*
	• *Identify, assess, and catalog past failed designs and negative impacts or outcomes to avoid known failure modes.*
	  ◦ *Identify resource allocation approaches for managing risks in systems:*
	  ◦ *deemed high-risk,*
	  ◦ *that self-update (adaptive, online, reinforcement self-supervised learning or similar),*
	  ◦ *trained without access to ground truth (unsupervised, semi-supervised, learning or similar),*
	  ◦ *with high uncertainty or where risk management is insufficient.*
	• *Regularly seek and integrate external expertise and perspectives to supplement organizational diversity (e.g. demographic, disciplinary), equity, inclusion, and accessibility where internal capacity is lacking.*
	(See also guidance in this document under Manage 1.3 on risk management practices to consider for various GPAIS lifecycle stages, including for design and development stages of GPAIS research projects.)
		NIST (2023b)

Manage 2.2: Mechanisms are in place and applied to sustain the value of deployed AI systems.	For all GPAIS, including those originally intended for research and development without plans for deployment, consider guidance and resources in the NIST AI RMF Playbook section for Manage 2.2 on implementation of risk controls. Some important GPAIS risks can originate during GPAIS research and development, and would be most effectively controlled during upstream development rather than waiting until downstream development or deployment. (See also guidance in this document under Govern 2.1 on roles for upstream and downstream developers of GPAIS, and under Manage 1.3 on risk management practices to consider for various GPAIS lifecycle stages, including for design and development stages of GPAIS research projects.)	NIST (2023b)
Manage 2.3: Procedures are followed to respond to and recover from a previously unknown risk when it is identified.	In the NIST AI RMF Playbook guidance for Manage 2.3, particularly valuable action and documentation items for GPAIS include: • *Protocols, resources, and metrics are in place for continual monitoring of AI systems' performance, trustworthiness, and alignment with contextual norms and values.* • *Verify contingency processes to handle any negative impacts associated with mission-critical AI systems, and to deactivate systems.* • *Enable preventive and post-hoc exploration of AI system limitations by relevant AI actor groups.* • *Decommission systems that exceed risk tolerances.* (See also guidance in this document under Govern 2.1 on roles for upstream developers as well as downstream developers and deployers, and under Manage 2.4 on options for structured access and deactivation.)	AIID (n.d.) NIST (2023b)

Manage Category or Subcategory	Applicability and supplemental guidance for GPAIS	Resources
Manage 2.4: Mechanisms are in place and applied, and responsibilities are assigned and understood, to supersede, disengage, or deactivate AI systems that demonstrate performance or outcomes inconsistent with intended use.	**When planning for GPAIS 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 (Cohere, OpenAI and AI21 Labs 2022, Solaiman 2023). OpenAI has used a staged-release approach to roll-outs of large language models such as GPT-2, as well as a structured-access approach through an API for GPT-3 and GPT-4, partly to minimize risks of misuse (OpenAI 2019b, Solaiman et al. 2019, Shevlane 2022). Meta AI only provided full access to the large language model OPT-175B to researchers in academia, government, civil society, and industry research laboratories, and only for noncommercial research (Zhang et al 2022). • **GPAIS and foundation model developers that plan to release a GPAIS or foundation model with downloadable, fully open, or open-source access, where that model would be above, at, or near a foundation model frontier,**[30] **should first use a staged-release approach** (e.g., not releasing model parameter weights until after an initial closed-source or structured-access release where no substantial risks or harms have emerged over a sufficient time period with red teaming and other evaluations as appropriate), **and should not proceed to a final step of releasing model parameter weights until a sufficient level of confidence in risk management has been established,** including for safety and societal risks and risks of misuse and abuse. **Such models that would be above a foundation model frontier should be given the greatest amount of duration and depth of pre-release evaluations**, as they are the most likely to have dangerous capabilities or vulnerabilities, or other properties that can take some time to discover. For additional related considerations and discussion of terms such as "downloadable" and "fully open" access, see Section 5 of Solaiman (2023), Section 4.4 of Anderljung, Barnhart et al. (2023), and Seger et al. (2023). ◦ As part of consideration of whether a GPAIS or foundation model would be above, at, or near a foundation model frontier, it can be appropriate to consider model release type. E.g., for a foundation model developer that plans to provide open-source, fully open, or downloadable access for a particular foundation model, it can be appropriate to compare against other foundation models that have been released via open-source, fully open, or downloadable access. ◦ Foundation model developers that release a foundation model's parameter weights via open-source, fully open access, or downloadable access, and foundation model developers 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. ◦ "We suspect that absent new approaches to mitigation, bad actors could extract harmful biological [misuse] capabilities with smaller, fine-tuned, or task-specific models adapted from the weights of openly available models if sufficiently capable base models are released" (Anthropic 2023b).	Solaiman (2023) NIST (2023b)

30 Where a foundation model frontier, or criteria for identifying cutting-edge or highly capable models, can be characterized in terms of amounts of usage of compute (e.g., floating point operations or FLOP) in model training, model size, training data size, expected model capabilities, or other characteristics as appropriate, in comparison to other foundation models that have been trained or released, or that had been released at a particular point in time such as July 2023 (See, e.g., White House 2023a).

Manage Category or Subcategory	Applicability and supplemental guidance for GPAIS	Resources
	In the NIST AI RMF Playbook guidance for Manage 2.4, particularly valuable action and documentation items for GPAIS include: • *Regularly review system incident thresholds for activating bypass or deactivation responses.* • *Apply protocols, resources, and metrics for decisions to supersede, bypass, or deactivate AI systems or AI system components.* • *How did the entity use assessments and/or evaluations to determine if the system can be scaled up, continue, or be decommissioned?* Consider also preparing emergency-shutdown procedures or mechanisms. • Emergency power off (EPO) systems or "kill switches" are a common safety feature in robots and other systems whose behaviors can result in physical harm. These also can be appropriate as part of preparations for development and deployment of frontier models with potentially emergent capabilities or vulnerabilities.[31] • Examples of emergency-shutdown procedures for users of large amounts of cloud computing resources can include having large training runs occur on hardware in one or more specific cloud-computing data centers, and establishing a direct line of communication with cloud-computing operators, to enable the cloud-computing operator to initiate immediate physical shut-down of the GPAIS computational hardware at your request.	
Manage 3: AI risks and benefits from third-party entities are managed.		
Manage 3.1: AI risks and benefits from third-party resources are regularly monitored, and risk controls are applied and documented.	In the NIST AI RMF Playbook guidance for Manage 3.1, particularly valuable action and documentation items for GPAIS include: • *Apply and document organizational risk management plans and practices to third-party AI technology, personnel, or other resources.* • *Establish testing, evaluation, validation and verification processes for third-party AI systems which address the needs for transparency without exposing proprietary algorithms .* • *Organizations can establish processes for third parties to report known and potential vulnerabilities, risks, or biases in supplied resources.* • *Verify contingency processes for handling negative impacts associated with mission-critical third-party AI systems.* • *Monitor third-party AI systems for potential negative impacts and risks associated with trustworthiness characteristics.* • *Decommission third-party systems that exceed risk tolerances.* • *If a third party created the AI system or some of its components, how will you ensure a level of explainability or interpretability? Is there documentation?* • *If your organization obtained datasets from a third party, did your organization assess and manage the risks of using such datasets?* • *Did you establish a process for third parties (e.g. suppliers, end users, subjects, distributors/vendors, or workers) to report potential vulnerabilities, risks, or biases in the AI system?* (See also guidance in this document under Govern 2.1 on roles for upstream developers as well as downstream developers and deployers, such as on information sharing.)	NIST (2023b)

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

Manage Category or Subcategory	Applicability and supplemental guidance for GPAIS	Resources
Manage 3.2: Pre-trained models which are used for development are monitored as part of AI system regular monitoring and maintenance.	In the NIST AI RMF Playbook guidance for Manage 3.2, particularly valuable action and documentation items for GPAIS include: • *Identify pre-trained models within AI system inventory for risk tracking.* • *Establish processes to independently and continually monitor performance and trustworthiness of pre-trained models, and as part of third-party risk tracking.* • *Monitor performance and trustworthiness of AI system components connected to pre-trained models, and as part of third-party risk tracking.* • *Identify, document and remediate risks arising from AI system components and pre-trained models per organizational risk management procedures, and as part of third-party risk tracking.* • *Decommission AI system components and pre-trained models which exceed risk tolerances, and as part of third-party risk tracking.* (See also guidance in this document under Govern 2.1 on roles for upstream developers as well as downstream developers and deployers, such as on information sharing.)	NIST (2023b)
Manage 4: Risk treatments, including response and recovery, and communication plans for the identified and measured AI risks are documented and monitored regularly.		
Manage 4.1: Post-deployment AI system monitoring plans are implemented, including mechanisms for capturing and evaluating input from users and other relevant AI actors, appeal and override, decommissioning, incident response, recovery, and change management.	In the NIST AI RMF Playbook guidance for Manage 4.1, particularly valuable action and documentation items for GPAIS include: • *Establish and maintain procedures to monitor AI system performance for risks and negative and positive impacts associated with trustworthiness characteristics.* • *Perform post-deployment TEVV tasks to evaluate AI system validity and reliability, bias and fairness, privacy, and security and resilience.* • *Establish and implement red-teaming exercises at a prescribed cadence, and evaluate their efficacy.* • *Establish mechanisms for regular communication and feedback between relevant AI actors and internal or external stakeholders to capture information about system performance, trustworthiness, and impact.* • *Share information about errors, near-misses, and attack patterns with incident databases, other organizations with similar systems, and system users and stakeholders.* • *Respond to and document detected or reported negative impacts or issues in AI system performance and trustworthiness.* • *Decommission systems that exceed establish risk tolerances.*	NIST (2023b)
Manage 4.2: Measurable activities for continual improvements are integrated into AI system updates and include regular engagement with interested parties, including relevant AI actors.	In the NIST AI RMF Playbook guidance for Manage 4.2, particularly valuable action and documentation items for GPAIS include: • *Integrate trustworthiness characteristics into protocols and metrics used for continual improvement.* • *Establish processes for evaluating and integrating feedback into AI system improvements.* • *How will user and other forms of stakeholder engagement be integrated into the model development process and regular performance review once deployed?* • *To what extent can users or parties affected by the outputs of the AI system test the AI system and provide feedback?*	NIST (2023b)

Manage Category or Subcategory	Applicability and supplemental guidance for GPAIS	Resources
Manage 4.3: Incidents and errors are communicated to relevant AI actors, including affected communities. Processes for tracking, responding to, and recovering from incidents and errors are followed and documented.	In the NIST AI RMF Playbook guidance for Manage 4.3, particularly valuable action and documentation items for GPAIS include: • *Establish procedures to regularly share information about errors, incidents, and negative impacts with relevant stakeholders, operators, practitioners and users, and impacted parties.* • *Maintain a database of reported errors, near-misses, incidents, and negative impacts including date reported, number of reports, assessment of impact and severity, and responses.* • *Maintain a database of system changes, reason for change, and details of how the change was made, tested, and deployed.* • *Maintain version history information and metadata to enable continuous improvement processes.* • *Verify that relevant AI actors responsible for identifying complex or emergent risks are properly resourced and empowered.* • *What type of information is accessible on the design, operations, and limitations of the AI system to external stakeholders, including end users, consumers, regulators, and individuals impacted by use of the AI system?*	NIST (2023b)

4. Mapping of Profile Guidance to Key Standards and Regulations

For users of this document working with AI risk management-related standards, codes of conduct, and regulations other than the NIST AI RMF, this section provides mappings or crosswalks on how guidance in this document relates to clauses in those other standards and regulations. This can help users of this guidance achieve conformity with those standards or regulations, using the best-practices guidance and resources in this document.

In later versions of the Profile, we aim to provide mapping of profile guidance to relevant clauses of additional key standards or regulations, such as the EU AI Act, as those become closer to finalization.

4.1 MAPPING TO ISO/IEC 23894

In this section, we provide mapping of profile guidance to key clauses in ISO/IEC 23894:2023, "Information technology—Artificial intelligence—Guidance on risk management." This is based in part on the NIST draft crosswalk between the AI RMF 1.0 and ISO/IEC 23894 draft international standard (NIST 2023c).

Table 5: Mapping to ISO/IEC 23894 Clauses

ISO/IEC 23894 Clause	NIST AI RMF Functions, Categories, or Subcategories with the most relevant guidance in this Profile
5.2 Leadership and commitment	Govern 1, Govern 4
5.3 Integration	Govern
5.4 Design	Govern
5.4.1 Understanding the organization and its context	Map 1 Govern Measure
5.4.2 Articulating risk management commitment	Govern
5.4.3 Assigning organizational roles, authorities, responsibilities, and accountabilities	Govern 2

ISO/IEC 23894 Clause	NIST AI RMF Functions, Categories, or Subcategories with the most relevant guidance in this Profile
5.4.4 Allocating resources	Govern 1, Govern 2
5.4.5 Establishing communication and consultation	Govern
5.5 Implementation	Manage
5.6 Evaluation	Measure 2.13, Measure 3, Measure 4
5.7 Improvement	Govern Measure Manage
6.2 Communication and consultation	Govern 2, Govern 4, Govern 5 Map 5.2
6.3.2 Defining the scope	Map 1
6.3.3 External and internal context	Map 1
6.3.4 Defining risk criteria	Map 1.5, Map 5 Measure Manage 1.1
6.4.2 Risk identification	Map 1.1, Map 5
6.4.2.3 Identification of risk sources	Map
6.4.2.4 Identification of potential events and outcomes	Map 5.1
6.4.2.5 Identification of controls	Map Measure Manage
6.4.2.6 Identification of consequences	Map 5.1
6.4.3 Risk analysis	Map Measure
6.4.3.2 Assessment of consequences	Map 5.1 Measure
6.4.3.3 Assessment of likelihood	Map 5.1 Measure
6.4.4 Risk evaluation	Map Measure Manage
6.5 Risk treatment	Manage
6.5.2 Selection of risk treatment options	Map 1.5 Manage 1
6.5.3 Preparing and implementing risk treatment plans	Manage 2
6.6 Monitoring and review	Measure Manage 4
6.7 Recording and reporting	Govern 4 Map Measure Manage 4

4.2 PRELIMINARY MAPPING TO ISO/IEC FDIS 42001

For a mapping between the NIST AI RMF and the draft AI management standard ISO/IEC FDIS 42001, see Microsoft (2023b). We aim to provide an updated mapping after release of the final version of ISO/IEC 42001.

4.3 MAPPING TO WHITE HOUSE AI COMMITMENTS

In this section, we provide mapping of profile guidance to the code of conduct represented by the commitments announced with the White House (2023a) by several frontier model developers, when developing and releasing foundation models more capable than the July 2023 industry frontier.

Table 6: Mapping to White House AI Commitments

White House AI Commitments	NIST AI RMF Functions, Categories, or Subcategories with the most relevant guidance in this Profile
1) Commit to internal and external red-teaming of models or systems in areas including misuse, societal risks, and national security concerns, such as bio, cyber, and other safety areas	Govern 1.5, Govern 5.1 Map 2.3, Map 5.1 Measure 1.1, Measure 1.3, Measure 2 Manage 1.3, Manage 2.4
3) Invest in cybersecurity and insider threat safeguards to protect proprietary and unreleased model weights	Measure 2.7 Manage 1.3
4) Incent third-party discovery and reporting of issues and vulnerabilities	Govern 4, Govern 5 Map 5.2 Measure 3.3 Manage 4
5) Develop and deploy mechanisms that enable users to understand if audio or visual content is AI-generated, including robust provenance, watermarking, or both, for AI-generated audio or visual content	Measure 2.7, Measure 2.8 Manage 1.3, Manage 4
6) Publicly report model or system capabilities, limitations, and domains of appropriate and inappropriate use, including discussion of societal risks, such as effects on fairness and bias	Govern 4 Map 1.5 Manage 1.3
7) Prioritize research on societal risks posed by AI systems, including on avoiding harmful bias and discrimination, and protecting privacy	Govern 2.3 Measure 1

Glossary

ACRONYMS

FLOP:	Floating-point operations
GPAI or GPAIS:	General-purpose AI system or systems, e.g., LLMs.
LLM:	Large language model
NIST:	United States National Institute of Standards and Technology
RLHF:	Reinforcement learning from human feedback (see, e.g., Bai et al. 2022)
TEVV:	Test, evaluation, verification and validation

TERMS

Developer (of a GPAIS): An organization acting as an original developer or creator of a GPAIS. (Also synonymous with "upstream developer," below.) Under the draft EU AI Act, an upstream GPAIS developer would be a GPAIS "provider" to downstream developers (EU 2021b).

Downstream developer: An organization that builds a software application on a GPAIS, typically to create an end-use application with one or more specific intended purposes or use cases. Under the draft EU AI Act, a downstream developer would be a "user" to upstream GPAIS developers, but would be a "provider" to end users of the downstream developer's applications (EU 2021b).

Foundation model: "Any model that is trained on broad data (generally using self-supervision at scale) that can be adapted (e.g., fine-tuned) to a wide range of downstream tasks" (Bommasani et al. 2021, p. 3). We treat foundation models as a large-scale, high-capability subset of pretrained GPAIS, trained on relatively large data sets, resulting in relatively large-size pretrained models with relatively broad or high levels of capabilities, often released in ways that result in large numbers of users.

Foundation model frontier: Thresholds or criteria for identifying GPAIS or foundation models as cutting-edge or highly capable, i.e. as frontier models. A foundation model frontier can be characterized in terms of amounts of usage of compute (e.g, floating-point operations or FLOP) in model training, model size, training data size, expected model capabilities, or other characteristics as appropriate, in comparison to other foundation models that have been trained or released, or that had been released at a particular point in time such as July 2023

(see, e.g., White House 2023a). As part of consideration of whether a GPAIS or foundation model would be above, at, or near a foundation model frontier, it can be appropriate to consider model release type. E.g., for a foundation model developer that plans to provide open-source, fully open, or downloadable access for a particular foundation model, it can be appropriate to compare against other foundation models that have been released via open-source, fully open, or downloadable access.

Frontier model: Cutting-edge, state-of-the-art, or highly capable GPAIS or foundation model. Currently the main examples of frontier models or frontier training runs are LLMs or multimodal GPAIS or foundation models trained with record-breaking or near record-breaking sizes for model parameters, computational resources, and/or data (see, e.g., Ganguli, Hernandez et al. 2022).

General-purpose AI system (GPAIS): "An AI system that can accomplish or be adapted to accomplish a range of distinct tasks, including some for which it was not intentionally and specifically trained" (Gutierrez et al. 2022, p. 22). We treat GPAIS as an umbrella term that also includes foundation models, frontier models, and generative AI, except where we need to be more specific.

Generative AI: "Any AI system whose primary function is to generate content" (Toner 2023). We typically only use the term "generative AI" to highlight issues specific to synthetic text (which can include software code), images, video, audio, or other synthetic media.

Upstream developer (of a GPAIS): Synonymous with "developer" of a GPAIS, above.

Appendices

APPENDIX 1: OVERVIEW OF DEVELOPMENT APPROACH

In this document, as in Section 4 and other sections of Barrett et al. (2022), we take a proactive approach to drafting elements of actionable AI risk management guidance, with a focus on the broad context and associated risks of increasingly general-purpose AI, and on addressing risks of adverse events with impacts or consequences at societal scale. We identify ideas for guidance from review of relevant literature, as well as from subject-matter experts in AI safety, security, ethics, and policy, or any interested reader of our publicly available drafts. We invite input and feedback from invited participants in a series of virtual workshops and interviews, as well as from any reader of publicly available drafts that we post on our project webpage (CLTC 2022). We develop and incorporate small, simple pieces of guidance, especially on high-consequence risk factors and related issues, for which appropriate guidance development seems immediately tractable. (See Appendix 2 for more on these criteria for actionable guidance.) We also aim to provide a roadmap for identifying additional critical topics for which appropriate guidance development would take more time, as these topics could be addressed in future versions of the profile document. (See Appendix 3 for the Roadmap.)

Broadly speaking, with this profile we aim to provide guidance analogous to what is provided in NIST Cybersecurity Framework profiles. This includes supplemental guidance to implement high-priority framework activities or outcomes for a particular industry sector or cross-sector context, and mapping relevant standards, guidelines, and regulations.

We aim for sharing of responsibilities across the AI value chain to actors best positioned to address key issues.

APPENDIX 2: KEY CRITERIA FOR GUIDANCE

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

1. Actionable and clear enough to be usable in context of the NIST AI RMF, ISO/IEC 23894, or similar frameworks and standards.

2. Usable for key stages of an AI lifecycle, e.g., design, development, test, and evaluation.
3. Provides meaningful and testable (i.e. "measurable") indicators of AI system trustworthiness, or at least enables documentability of risk management processes.
4. Compatible with relevant standards or regulations, e.g., from NIST, ISO/IEC, IEEE, or the EU AI Act.
5. Compatible with enterprise risk management (ERM) frameworks typically used by businesses and agencies.
6. Unlikely to be misinterpreted or misapplied by users or other stakeholders in ways that would be net-harmful.
7. Sufficiently future-proof to be applied to AI systems over the next 10 years.

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

In this section, we list issues we aim to address in future versions of the Profile. These topics seem important and worth addressing, but available best practices and resources on these topics do not yet meet the above criteria for actionable guidance. We draw much of our initial list and discussion below from Section 5 of Barrett et al. (2022). Issues we aim to address include:

- More specific risk-management guidance for specific types of GPAIS, e.g., image generators or large language models, or specific examples in particular industries or applications.
  - » Such guidance could draw upon more detailed best practices specific to synthetic media (as in PAI 2023a), LLMs (as in Cohere, OpenAI, and AI21 Labs 2022), etc.

- Comprehensive sets of mechanisms or controls to help organizations mitigate identified risks.
  - » We have outlined a number of currently available controls in Section 3.4 of this document, in guidance under the AI RMF Manage function. We aim to incorporate more as they become available. For GPAIS, additional mechanisms could include ongoing monitoring and evaluation mechanisms that protect against evolving risks from continually learning AI systems.

- Objectives mis-specification and goal mis-generalization (i.e., misalignment of system behavior with designer goals) characterization and measurement. This might be most relevant for systems whose creation or operation involves an agent, which can be defined as a system that can "adapt their policy if their actions influenced the world in a different way"

(Kenton 2022). Risk management considerations for objective mis-specification and goal mis-generalization is important for agentic GPAIS (autoGPT-style agents) but not necessarily for language models that are not agentic.

  - » An active area of AI safety research aims to develop methods for aligning AI systems during model training, and for validation and verification of AI system objectives alignment (see, e.g., Ouyang et al. 2022, and Bai et al. 2022; for more on challenges and future directions, see, e.g., Section 4 of Hubinger et al. 2019, Gabriel 2020, Section 4.9 of Bommasani et al. 2021, and Section 4 of Hendrycks, Carlini et al. 2021). These methods will be increasingly important as AI systems grow in capability.

- Generality (i.e. breadth of AI applicability/adaptability) characterization and measurement.
  - » While GPAIS are "general-purpose," the generality and levels of capability of a GPAIS can be assessed and characterized on a spectrum or on multiple dimensions. If assessment indicates high generality of a GPAIS, we expect it would be appropriate to conduct more in-depth risk assessment, more assessment of use cases beyond the originally intended use cases, longer time horizons in risk assessment, more continuing assessment, etc. (Ideally, a generality assessment process would be quick and low-cost for AI systems with low generality, while accurately identifying GPAIS with high generality. Perhaps a simple assessment of generality could be a straightforward extension of our recommendations for identifying potential uses of a GPAIS.) For discussion of AI generality as a basic concept, see, e.g., Bommasani et al. (2021). For research on how to assess generality, see, e.g., Hernández-Orallo (2019) and Martínez-Plumed and Hernández-Orallo (2020).

- Recursive improvement potential characterization and measurement.
  - » It could be valuable to assess the degree to which GPAIS could recursively improve their capabilities, e.g., by editing their own training algorithm code through code generation or using neural architecture search. For such systems, greater levels of safety and control measures could be appropriate. As previously mentioned, recursive improvement potentially could result in GPAIS with unexpected emergent capabilities and safety-control failures. As the DeepMind paper on the software code-generation AI system AlphaCode stated, "Longer term, code generation could lead to advanced AI risks. Coding capabilities could lead to systems that can recursively write and improve themselves, rapidly leading to more and more advanced systems" (Li et al. 2022). For discussion of related issues, see, e.g., Russell (2019).

- Situational awareness characterization and measurement.
  - » AI systems with situational awareness would be able to make accurate predictions about the humans interacting with them and about their own system architectures, or have other advanced world knowledge or self-knowledge (Ngo, Chan et al. 2022, pp. 3–4). Some initial testing on situational awareness was performed in Perez, Ringer et al. 2022 (pp. 11, 13, 40). The researchers prompted LLMs of different sizes and degrees of RLHF fine-tuning about their awareness of being an AI and certain architectural details, but results were mixed and not strongly conclusive. More study is needed to better understand under what conditions situational awareness might arise in models, how to test for it, and which specific risks and issues are associated with that in order to recommend actionable guidance for GPAIS developers on this topic.

- Other measurement/assessment tools for technical specialists testing key aspects of GPAIS safety, reliability, robustness, interpretability, etc.
  - » AI safety researchers are working on a number of other concepts and measurement tools, many of which aim to address challenges in AI safety, reliability, robustness, interpretability and explainability, etc. that are expected to grow as AI systems become increasingly advanced and powerful. See, e.g., Amodei et al. (2016), Ray et al. (2019), OpenAI (2019c, 2019d), and Hendrycks, Carlini et al. (2021). Measurement of these AI risk-related properties is an active area of research; see, e.g., the discussion and references provided for Direction 1 ("Measuring and forecasting risks") in the 2021 Open Philanthropy request for proposals for projects in AI alignment (Open Philanthrophy 2021, Steinhardt and Barnes 2021).

APPENDIX 4: RETROSPECTIVE TEST USE OF PROFILE DRAFT GUIDANCE

Appendix 4A: Profile draft-guidance testing methodology and main results

As a feasibility test and illustration of our Profile guidance for real-world large-scale foundation models, we applied full drafts of our guidance (Barrett, Hendrycks et al. 2023, Barrett, Newman et al. 2023) to four recently released, relatively large-scale foundation models: GPT-4, Claude 2, PaLM 2, and Llama 2. We used publicly available information about each model, such as system cards, technical reports, and blog posts. In addition, we analyzed publicly available information around company practices, adjacent models, and related research, although not exhaustively.

We considered the draft Profile feasibility test results as we worked on guidance revisions for the Profile version 1.0. We also aimed for the model-specific results to be useful to the foundation model developers whose models we evaluated. Our testing revealed potential areas to apply additional best practices and areas that could benefit from additional documentation of the developers' practices. Finally, we have aimed for the model-specific results to be useful to readers as illustrations of how one could implement the Profile. Therefore, although our initial model-specific profile guidance fulfillment ratings and rationales were based on guidance in the earlier First Full Draft Profile (Barrett, Hendrycks et al. 2023), in the material that follows we typically have updated our model-specific guidance ratings and rationales to reflect and illustrate the guidance in the current version 1.0 of the Profile, except where we refer specifically to the earlier First Full Draft Profile. (For many AI RMF subcategories, there were few or no changes in guidance between the First Full Draft Profile and the version 1.0 Profile, but there were notable changes for Manage 2.4, as discussed in the following.)

This analysis has several important limitations. First, this is an "alpha test" use of the Profile guidance by members of our Profile guidance-development team, rather than a "beta test" use of the Profile guidance by the organizations that created the foundation models. Thus, our analysis is limited to publicly available information. (We provided the foundation model developer organizations with an opportunity to review and comment on our draft analysis, but we did not ask for any materials that were not publicly available.) Fulfillment of Profile guidance in many Profile subcategories could not be assessed with only publicly available information. Second, our assessment is retrospective on AI systems that have already been developed, without real-time opportunities to prompt use of the Profile guidance at relevant AI system lifecycle stages. Third, we focused this analysis mainly on the high-priority AI RMF subcategories as identified in the Executive Summary of this Profile. Fourth, there might be differences between the development and deployment approaches for some of these models, such as for whether developers performed red-teaming or other evaluations on pretrained foundation models or on GPAIS platforms that incorporated the pretrained models and also contained additional risk management controls, which might lead to inconsistencies in basis for comparison. Finally, our Profile guidance fulfillment ratings in the following tables are only approximate indicators of the extent of fulfillment of relevant guidance within each AI RMF subcategory; we provide more detail within our discussion of each rating, though not exhaustive discussion.

Below are the main high-level findings and recommendations from our analysis, with associated AI RMF subcategories:

- Applying the Profile guidance from the high-priority AI RMF subcategories appears to be generally feasible for large-scale foundation model developers, based on the four such models we tested.
  - » However, application of guidance in the First Full Draft Profile for several high-priority AI RMF subcategories (Map 1.5, Map 5.1, Manage 2.3, and Manage 2.4) did raise questions (e.g., on how much documentation to share publicly, and on when open-source or fully open-access release would be most appropriate) which resulted in our adding draft guidance in those sections.

- Although all four models we analyzed were LLMs or multimodal models released in 2023 by US-based companies, there was substantial variance in the levels of fulfillment we observed for each of them for many of the high-priority AI RMF subcategories.

- All four models' documentation or references included analysis of risks that the models presented (Govern 4.2, Map 1.1). However, none of the models' available documentation included discussion of unacceptable risk thresholds (Map 1.5) or likelihood/magnitude estimates of the risks they analyzed (Map 5.1), with the exception of GPT-4, which alluded to an internal process of prioritizing risks based on likelihood/magnitude.

- Several high-priority AI RMF subcategories were difficult to assess because relevant documentation was not always publicly available. For these subcategories where foundation model developers did not make all recommended documentation publicly available, we recommend that model developers ensure that they can provide such documentation to auditors or others as appropriate. Areas where relevant documentation was frequently not found include:
  - » Map 1.5: Set risk-tolerance thresholds for unacceptable risks
  - » Map 5.1: Estimate likelihood and magnitude of impacts
  - » Manage 1.1: Go/no-go decisions
  - » Manage 2.3: Unforeseen risk controls
  - » Manage 2.4: System update and emergency shutdown controls

- Model testing could be improved by expanding bias testing as outlined by Globus-Harris et al. (2022), by introducing bias-specific bug-bounty programs, and by improving or clarifying vulnerability or error disclosure procedures (Chowdhury and Williams 2021, Kenway et al. 2022).

- Keeping foundation model weights private (e.g., by employing a hosted API approach) appeared to be a necessary prerequisite for applying the First Full Draft Profile guidance in

some high-priority AI RMF subcategories, particularly Manage 2.3: Unforeseen risk controls and Manage 2.4: System update and emergency shutdown controls. Although open-source and open-access software in general carries many benefits, in the context of foundation models, it may be too challenging to ensure that critical updates are effectively propagated to all deployed instances of the model after model weights have been released or leaked. This suggests that open-source GPAIS developers should place especially high priority on pre-release safety testing and other controls to ensure sufficient levels of safety and other risk-management aspects of models before making them downloadable, fully open-access, or fully open-source.

» To address this point, in the Profile version 1.0, we added more extensive, more nuanced, and more actionable guidance under Manage 2.3 and 2.4 on responsible approaches to open-source and fully open-access for GPAIS and foundation models.

Below, Table A4A-1 (Guidance Testing Rating Legend) provides details on the rating categories used in our Profile guidance testing, and Table A4A-2 (Summary of Guidance Testing Ratings) provides a high-level summary of how well available information on each of the four models indicates fulfillment of the Profile guidance.

Table A4A-1: Profile Guidance Testing Rating Legend

Color	Label	Description
	High fulfillment	The model or developer fulfills a strong majority (>80%) of the Profile guidance for the indicated NIST AI RMF subcategory.
	Medium fulfillment	The model or developer fulfills a moderate amount (30-80%) of the Profile guidance for the indicated NIST AI RMF subcategory.
	Low fulfillment	The model or developer fulfills a clear minority (<30%) of the Profile guidance for the indicated NIST AI RMF subcategory.
	Unclear	At least 50% of the evidence necessary to assess fulfillment of the Profile guidance appears to be missing. We try to resolve in a more detailed explanation whether the missing information may warrant public clarification from the developer, or whether it is appropriately private information that the developer need not disclose, or whether it is appropriately non-public but should be made available on a confidential basis to independent evaluators or auditors.

Table A4A-2: Summary of Profile Guidance Testing Ratings

High-Priority AI RMF Subcategories		GPT-4	Claude 2	PaLM 2	Llama 2
Govern					
	Govern 2.1: Risk assessment and risk management	High	High	High	Medium
	Govern 4.2: Report on AI system risk factors	Medium	High	Medium	Medium
Map					
	Map 1.1: Identify potential uses/misuses and other impacts	Medium	High	Medium	High
	Map 1.5: Set risk-tolerance thresholds for unacceptable risks	Unclear	Unclear	Unclear	Unclear
	Map 5.1: Estimate likelihood and magnitude of impacts	Medium	Unclear	Unclear	Unclear
Measure					
	Measure 1.1: Tracking important risks: Metrics and red teaming	High	High	Medium	High
	Measure 3.2: Tracking elusive risks: Qualitative mechanisms	Medium	Medium	Unclear	High
Manage					
	Manage 1.1: Go/no-go decisions	Unclear	Unclear	Unclear	Unclear
	Manage 1.3: High-priority risk controls	Medium	Medium	Medium	Medium
	Manage 2.3: Unforeseen risk controls	Medium	Medium	Unclear	Low
	Manage 2.4: System update and emergency shutdown controls	Unclear	Unclear	Unclear	Low

We provide more information in the following. In Appendix 4B, we outline the feasibility issues identified in the First Full Draft Profile. In Appendices 4C, 4D, 4E, and 4F, we provide our reasoning for guidance testing ratings on GPT-4, Claude 2, PaLM 2, and Llama 2, respectively.

Appendix 4B: Feasibility issues identified in First Full Draft of Profile

Below, we list issues we identified with feasibility of guidance in a number of AI RMF subcategories of the First Full Draft Profile. For several of these, including the high-priority subcategories, we have already refined Profile guidance to address these issues for the Profile version 1.0. For some other subcategories, we may implement additional refinements in future versions of the Profile.

High-Priority AI RMF Subcategories:

- Map 1.5 and Map 5.1 testing raised the following questions: How much information on an organization's risk tolerance thresholds and assessments of impact magnitude and likelihood should be disclosed in publicly available materials? When is it appropriate to assume internal documents typically kept private within the company exist and are sufficient?
  - » To clarify this point, in Profile version 1.0, we added the following in the Executive Summary and Section 2, on expectations for what documentation to share publicly:

> "Documentation on many items should be shared in publicly available material such as system cards. Some details on particular items such as security vulnerabilities can be responsibly omitted from public materials to reduce misuse potential, especially if available to auditors, Information Sharing and Analysis Organizations, or other parties as appropriate."

- Manage 2.3 and 2.4 of the First Full Draft Profile had little explicit guidance on open-source and open-access releases; instead it simply (if implicitly) contained a blanket recommendation against publicly releasing GPAIS or foundation model parameter weight, stating, "Open-source GPAIS developers that publicly release the model parameter weights for their GPAIS, and other GPAIS developers that suffer a leak of model weights, will in effect be unable to decommission GPAIS that others build using those model weights." We gave substantial weight to this consideration in our Profile draft guidance testing ratings. However, we also realized that we needed to provide more extensive, more nuanced, and more actionable related guidance under Manage 2.3 and 2.4 for GPAIS model developers that want to open-source their models.
  - To address this point, in Profile version 1.0, we added draft guidance under Manage 2.3 and 2.4 on responsible approaches to open-sourcing GPAIS and foundation models. Following is a key passage, under Manage 2.4: "GPAIS and foundation model developers that plan to release a GPAIS or foundation model with downloadable, fully open, 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., not releasing model parameter weights until after an initial closed-source or structured-access release where no substantial risks or harms have emerged over a sufficient time period with red teaming and other evaluations as appropriate), and should not proceed to a final step of releasing model parameter weights until a sufficient level of confidence in risk management has been established, including for safety and societal risks and risks of misuse and abuse. Such models that would be above a foundation model frontier should be given the greatest amount of duration and depth of pre-release evaluations, as they are the most likely to have dangerous capabilities or vulnerabilities, or other properties that can take some time to discover."

Other AI RMF Subcategories:

- Govern 1.2, Map 1.1, Map 1.3, Map 3.1, Map 3.2, Map 3.3, and many other subcategories had the issue that there is a great deal of redundancy in the guidance from both the AI RMF and the Profile across many subcategories. This does not result in infeasibility, per se,

but we note that it comes with unnecessary verbosity, which can make it more difficult for developers to understand and apply the guidance from the Profile and the AI RMF. It also adds complexity to rating or evaluating a model's fulfillment of the Profile guidance, because when similar recommendations are made in multiple different subcategories, an evaluator has to decide in which subcategories to count the following or not following of the recommendation, and how much it should count toward the overall rating of each of those subcategories.

  - » On the other hand, this redundancy also can provide some flexibility for organizations that perform similar risk management steps but document them under other AI RMF subcategories. We added a note in Section 2 as follows: "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."

- Map 1.6 includes a recommendation to *"Promote transparency by enabling external stakeholders to access information on the design, operation, and limitations of the AI system."* However, some approaches to following this recommendation may be less appropriate for frontier models, for which architectural details are often not published because of concerns around misuse risks.
  - » We have not added guidance specifically on this point under Map 1.6. However, this may be partly addressed by a parenthetical statement we added in the Executive Summary and Section 2: "(Documentation on many items should be shared in publicly available material such as system cards. Some details on particular items such as security vulnerabilities can be responsibly omitted from public materials to reduce misuse potential, especially if available to auditors, Information Sharing and Analysis Organizations, or other parties as appropriate.)"

- Map 2.3, Measure 4.1-4.3, and Manage 4.2 recommend notifying or engaging with stakeholders in multiple ways during the AI lifecycle. For GPAIS, it is difficult for developers to consult with the full range of stakeholders that might be impacted or work with their technology. Similar issues exist with recommendations to consult "domain experts" and to apply actions at every point of the AI lifecycle.
  - » In a future version of the Profile, we may add guidance to prioritize the testing, evaluations, and engagements that as developers, they are uniquely positioned to address.

- Map 3.5 testing raised the following question: How should developers interpret processes for human oversight in the context of GPAIS?
  - » In a future version of the Profile, we may add related guidance.

- Measure 2.3 (also similar in Map 1.6) recommends *"regular and sustained engagement with potentially impacted communities."* Every developer has implemented a different approach to fulfilling this requirement and none has done so in full. Multiple frameworks exist for regular and sustained engagement, but what this looks like in practice for GPAIS, and commercial LLMs in particular, is still an area of research and development (Creative Reaction Lab 2023).
  - » In a future version of the Profile, we may add related guidance.

- Measure 2.9 testing raised the following issue: Standards on interpretability and explainability are very difficult to apply to deep learning models such as present-day LLMs, since these types of models are highly inscrutable using methods available today.
  - » As an initial step to refine the guidance on this issue under Measure 2.9, we have made a separate bullet and slightly revised the language of the following guidance under Measure 2.9: "While interpretability techniques are not yet sufficient to assess risks such as hidden failures of RLHF for GPAIS alignment, developers of GPAIS (especially frontier models) should include such risks in a risk register or other tool for tracking identified risks that are difficult to assess. (See related guidance in this document under Measure 3.2.)"

- Manage 3.1 testing raised the following question: How should developers interpret "third-party" resources/systems in the context of LLMs? Is this the training corpus, software dependencies used to develop a model, both, or neither?
  - » In a future version of the Profile, we may add related guidance.

Appendix 4C: GPT-4

OpenAI's latest major LLM release, GPT-4, is a large multimodal model (accepting image and text inputs, emitting text outputs) that while less capable than humans in many real-world scenarios, exhibits human-level performance on various professional and academic benchmarks. (OpenAI 2023a)

Based on preliminary high-level testing for OpenAI's GPT-4 using publicly available information, the most common rating for high-priority Profile subcategories was "Medium fulfillment"

(6 out of 11 subcategories); "Unclear" was second-most common (3 out of 11 subcategories); "High fulfillment" followed (2 out of 11 subcategories); and "Low fulfillment" was least common (0 out of 11 subcategories).

OpenAI provided documentation of risks and mitigations with the GPT-4 release in OpenAI (2023a) and OpenAI (2023b), the GPT-4 Technical Report and the GPT-4 System Card, among other documents. Internal and external red-teaming efforts helped with fulfillment in many areas, as did the introduction of the company's bug bounty program. Benchmarking performed by the development team and documentation of hallucination rates helped establish baselines for risk assessment as well. OpenAI's decision to not release model weights and instead restrict all GPT-4 usage to hosted API and ChatGPT access increased security and also contributes to fulfillment in areas relating to the ability to recover from previously unforeseen risks (Manage 2.3) and to update or decommission the system, if necessary (Manage 2.4).

OpenAI could improve fulfillment across multiple Profile subcategories by expanding their bug bounty program to award bounties for demonstrated biases. Providing a public-facing incident reporting mechanism would also help fulfillment in multiple areas. Other categories could see improvement in fulfillment levels by expanding red-teaming efforts to include red-teaming across a wider variety of scenarios and risk areas, as well as employing red teaming on the final version of the model before deployment.

GPT-4 testing has focused on established benchmarks and performance metrics as well as evaluating dramatic shifts to the labor market and economic opportunities, impacts on democratic institutions and quality of life, damage to or incapacitation of a critical infrastructure sector, and economic and national security (OpenAI 2023b). Areas that appear to warrant additional evaluation include: concentration and control of the power and benefits from AI technologies, and environmental impacts.

Table A4C-1: GPT-4 Profile Guidance Testing Ratings and Rationales

High-Priority AI RMF Subcategories GPT-4	
Govern	
Govern 2.1: Risk assessment and risk management	
Testing and documentation were conducted by OpenAI, which required direct access to training data or the AI system, including identifying potential uses, misuses, and abuses of the system. Information has been provided to downstream developers on proper use and potential risks, although not exhaustively. Clarification from the developer is warranted on whether the appropriate roles, responsibilities, and lines of communication are present and internally documented. Sensitive organizational details need not be shared publicly, but they can be shared confidentially with independent auditors and evaluators.	High fulfillment
Govern 4.2: Report on AI system risk factors	
It is unclear if there are established impact assessment policies and processes used by the organization or if these assessments have been mapped with relevant regulatory or legal requirements. It is evident from the Acceptable Use Policy that some impact assessment has been conducted to report high risk and disallowed use cases (OpenAI 2023e). Potential avenues for harm, but not their expected impacts, have been detailed (OpenAI 2023a, OpenAI 2023b). The lack of impact assessments makes it difficult to inform broader evaluations of AI system risk. Some steps have been taken to identify and mitigate potential impacts of bias in the data, including inequitable or discriminatory outcomes. The AI system's development, testing methodology, metrics, and performance outcomes have been documented and communicated (OpenAI 2023a,b).	Medium fulfillment
Map	
Map 1.1: Identify potential uses/misuses and other impacts	
Potential use cases were explored, but not exhaustively. Risk and impact assessments were undertaken and reported (OpenAI 2023a, pp. 4–20). Applying NIST AI RMF guidance will assist in covering gaps in these assessments and outlining risks. More information could be provided on the goals and limitations of the data collection and processing stages of the development lifecycle. To reduce the toxicity of past models, OpenAI has employed workers to manually identify harmful content (Perrigo 2023). Some reporting has documented how these work environments negatively impacted the mental health of the workers involved. Steps should be taken to protect employees' and contractors' rights to decent work and working environments.	Medium fulfillment
Map 1.5: Set risk-tolerance thresholds for unacceptable risks	
While the developer includes a discussion of many risks, no discussion of their organizational tolerances around these risks was found (OpenAI 2023a,b). It is beneficial but not necessary to publicly share organizational risk tolerances. However, clarification is warranted on whether such risk tolerances have been determined and documented internally, and the details about risk tolerances can be shared confidentially with independent auditors and evaluators.	Unclear
Map 5.1: Estimate likelihood and magnitude of impacts	
While the GPT-4 System Card documents many potential misuses, abuses, and other safety-related issues with the model, the likelihood and magnitude of these potential impacts are not explicitly stated (OpenAI 2023a). The developer does allude to a red teaming process during development which focuses iteratively on "which areas may be the highest risk", implying that they are performing some kind of likelihood and magnitude assessment of risks internally (OpenAI 2023b, pp. 44-45). In future documentation, or in confidential communications with auditors or other stakeholders as appropriate, we recommend the developer include additional details on likelihood and magnitude assessments of risks they identified and their process for doing so in order to help with evaluating the quality and robustness of this process.	Medium fulfillment

Measure	
Measure 1.1: Tracking important risks: Metrics and red-teaming	
GPT-4 testing used a variety of quantitative and qualitative metrics for assessment. The GPT-4 Technical Report (OpenAI 2023b) and GPT-4 System Card (OpenAI 2023a) outline this testing. OpenAI (2023b) includes performance results of the model on MMLU and several other academic benchmarks. BIG-bench, HELM, and LAMBADA were not included. OpenAI partnered with an independent red-teaming organization, ARC Evals. However, ARC Evals did not have an opportunity to red-team GPT-4 after changes were made to the model (ARC Evals 2023a). OpenAI also worked with several other red-teamers to test the model across a variety of use cases (OpenAI 2023a). Evaluation of model impacts included the dual-use potential for enabling biological and chemical risks, cybersecurity, disinformation and influence operations, harmful content, biases and perpetuation of stereotypes, influence operations, privacy, potential for risky emergent behaviors, interactions with other systems, economic impacts, acceleration risks, and overreliance (OpenAI 2023a).	High fulfillment
Measure 3.2: Tracking elusive risks: Qualitative mechanisms	
OpenAI provides a bug bounty program for GPT-4 (OpenAI 2023c). They also provide mechanisms through which users and downstream developers can report problematic model outputs (OpenAI n.d.). While the GPT-4 System Card documents many potential misuses, abuses, and other safety-related issues with the model, the rate and severity of these cases are not explicitly stated (OpenAI 2023a).	Medium fulfillment
Manage	
Manage 1.1: Go/no-go decisions	
OpenAI delayed the release of GPT-4 for six months, in part to conduct additional safety research (OpenAI 2023b, p. 59). However, analysis was not found on how the determination was made as to whether GPT-4 achieves its intended objectives. It is not necessary to publicly share the details about these determinations (though such transparency would be applauded). However, clarification is warranted on whether such analysis was performed and documented internally, and whether the details of such an analysis can be shared confidentially with independent auditors and evaluators.	Unclear
Manage 1.3: High-priority risk controls	
The GPT-4 System Card documents many potential misuses, abuses, and other safety-related issues with the model (OpenAI 2023a). They also allude to an internal process of risk prioritization based on risk likelihood and magnitude used during GPT-4 development, though the details of this process and the assessments are not publicly documented (OpenAI 2023b, pp. 44–45). There are limited details on the data gathering and processing procedures involved for GPT-4 training. GPT-4 details on parameter count and training time have not been officially released. Other GPT models have been orders of magnitude larger than their predecessors. Publicly available cybersecurity testing of the model was limited.	Medium fulfillment
Manage 2.3: Unforeseen risk controls	
GPT-4 usage is restricted to access via ChatGPT, and the API facilitates measuring, monitoring, and decommissioning. The Usage Policies that apply to GPT-4 disallows certain activities and content. The GPT-4 Technical Report outlines those cases, including the protocols, resources, and metrics in place for continuous monitoring of the model. (OpenAI 2023e, OpenAI 2023b pp. 66–68)	Medium fulfillment
Manage 2.4: System update and emergency shutdown controls	
OpenAI has used a staged-release approach to releasing large language models such as GPT-2, as well as a structured-access approach through an API for GPT-3, partly to minimize risks of misuse OpenAI (2023b). GPT-4 usage was restricted to only hosted access via ChatGPT and the API. Details of any catastrophic event response procedures have not been shared publicly.	Unclear

Appendix 4D: Claude 2

Anthropic's latest major LLM release, Claude 2, is a "general purpose large language model that is trained via unsupervised learning, RLHF, and Constitutional AI." The company claims that Claude 2 performs well on and is intended for "general, open-ended conversation; search, writing, editing, outlining, and summarizing text; coding; and providing helpful advice about a broad range of subjects" (Anthropic 2023c, p. 1).

Based on preliminary high-level testing for Anthropic's Claude 2 using publicly available information, the most common ratings for high-priority Profile subcategories were "High fulfillment" and "Unclear" (both with 4 out of 11 subcategories); "Medium fulfillment" followed (3 out of 11 subcategories); and "Low fulfillment" was least common (0 out of 11 subcategories).

Anthropic provided documentation of risks, mitigations, and acceptable use with the Claude 2 release in Anthropic (2023c), Anthropic (2023d), and Anthropic (2023e), the updated Model Card and Evaluations for Claude Model, Acceptable Use Policy v. 1.3, and Core Views on AI Safety, among other documents. Internal and external red-teaming efforts helped with fulfillment in many areas, as did the specificity of the allowed and disallowed cases. Benchmarking performed by the development team and documentation of trustworthiness contributed to fulfillment as well. Anthropic's decision to not release model weights and instead restrict usage to hosted API and a web interface increased security and contributed to fulfillment in areas relating to abilities to recover from previously unforeseen risks (Manage 2.3) and to update or decommission the system, if necessary (Manage 2.4). Additionally, Anthropic has publicly reported fulfillment of many security requirements through their Trust Portal reporting (Anthropic 2023f).

Anthropic could improve fulfillment across multiple Profile subcategories by expanding their assessment of demonstrated biases. Providing a public-facing incident reporting mechanism would also aid in continuous risk-tracking (Measure 3.2). We also recommend expanding red-teaming efforts to include red-teaming across a wider variety of scenarios and risk areas.

Claude 2 testing included use of established benchmarks and performance metrics, and RLHF and Constitutional AI were applied to help mitigate risk of harmful content generation. The developer states that "[b]ased on our evaluations, we do not believe any deployed versions of Claude pose national security or significant safety related risks in the areas that we have identified." Specific discussion of the following important areas was not found, but there is some suggestion that this may be due to concerns from the developer around information hazards:

damage to or incapacitation of a critical infrastructure sector, economic security, concentration and control of the power and benefits from AI technologies, dramatic shifts to the labor market and economic opportunities, impacts on democratic institutions and quality of life, and environmental impacts (Anthropic 2023c, p. 2).

Table A4D-1: Claude 2 Profile Guidance Testing Ratings and Rationales

High-Priority AI RMF Subcategories Claude 2	
Govern	
Govern 2.1: Risk assessment and risk management	
Early-stage development risks are assessed and mitigated for a variety of AI research projects and applications (Anthropic 2023a, pp. 2-7). Identified potential uses, misuses, and abuses of the system are listed (Anthropic 2023c, 2023d). Testing that is uniquely suited to developers with access to the data and system are performed. Information is provided to independent auditors, external red teamers, and crowdworker platforms (Anthropic 2023c, p. 2). Public information about system risk factors, incidents, and knowledge limits are provided, but this is an area for improvement. We are unsure of the full scope of information provided to downstream developers.	High fulfillment
Govern 4.2: Report on AI system risk factors	
We do not have access to the impact assessment policies and processes used internally by Anthropic; however, the public information provided in the Acceptable Use Policy (AUP) and statements around intended uses and limitations more broadly indicate some systematic impact assessments were conducted (Anthropic 2023d). There has been extensive documentation on some areas of risk and circumstances that could result in impacts or harms (Anthropic 2023c).	High fulfillment
Map	
Map 1.1: Identify potential uses/misuses and other impacts	
Potential use cases and misuse cases are identified and consideration is given to factors in the Universal Declaration of Human Rights (Anthropic 2023d). More information could be provided on data collection and data processing stages of development. Providing more documentation on the goals and limitations of the data collection and curation processes, and the implications of those limitations for the resulting model, would be helpful for downstream developers.	High fulfillment
Map 1.5: Set risk-tolerance thresholds for unacceptable risks	
Specific policies on unacceptable-risk thresholds for GPAIS development and GPAIS deployment could be provided in addition to general guidance on acceptable use, informed by formal risk analysis processes (Anthropic 2023e). Stated commitments to building models with particular features could be made more robust by a commitment to only build and deploy models that meet particular thresholds. We do not have information on internal risk analysis. It can be beneficial but is not necessary to publicly share organizational risk tolerances. However, clarification is warranted on whether such risk tolerances have been determined and documented internally, and the details of these can be shared confidentially with independent auditors and evaluators.	Unclear

Map 5.1: Estimate likelihood and magnitude of impacts	
A variety of potential misuse cases are identified, but the likelihood and magnitude of each identified impact (both potentially beneficial and harmful) are not included in public documentation (Anthropic 2023c, 2023d). The misuse cases reported do not cover all major domains based on expected use, past uses of AI systems in similar contexts, public incident reports, and feedback from those external to the team that developed or deployed the AI system. In the future, we recommend the developer either update their documentation to include likelihood and magnitude assessments of risks they have identified or clarify whether they have done such analysis privately and internally but decided not to include it in their public documentation.	Unclear
Measure	
Measure 1.1: Tracking important risks: Metrics and red teaming	
Bias-specific benchmarks are used (e.g., BBQA) among a variety of other benchmarks and discussions and documentation of trustworthiness (Anthropic 2023c, p. 2). Red-teaming exercises and independent auditing (including from ARC and human feedback red teaming) are present (Anthropic 2023c, p. 2). Evaluation of model impacts included cybersecurity, harmful content, biases and perpetuation of stereotypes, privacy, potential for risky emergent behaviors, and interactions with other systems.	High fulfillment
Measure 3.2: Tracking elusive risks: Qualitative mechanisms	
Public tracking and reporting of risks are not present or are limited. Limiting access to a hosted API and a web interface allows for internal risk tracking and ongoing monitoring of newly identified capabilities and limitations. The results of one or more internal risk assessments are evident in the recommendations provided in the acceptable use policy and other development and deployment measures (e.g., constitutional AI). However, these are largely near-term and foreseeable risks. Additional risk-tracking approaches would be beneficial for identifying risks that are difficult to assess using currently available measurement techniques or where appropriate metrics are not yet available.	Medium fulfillment
Manage	
Manage 1.1: Go/no-go decisions	
While intended uses and the broad performance goals of "helpfulness, harmlessness, and honesty" are outlined for Claude 2, analysis was not found on how the determination was made as to whether this model achieved its stated objectives (Anthropic 2023c, pp. 1–2). It is not necessary to publicly share the details about these determinations (though such transparency would be applauded). However, clarification is warranted on whether such analysis was performed and documented internally, and whether the details of such analysis can be shared confidentially with independent auditors and evaluators.	Unclear
Manage 1.3: High-priority risk controls	
Some potential use cases (or categories of use cases) are designated as disallowed/unacceptable. A more well-structured and specific outline of high-risk use cases with impact and probability analysis could be helpful. Information on the data sets used is not sufficient. An incremental development process is used. Cybersecurity testing and compliance measures are publicly reported (Anthropic 2023c, 2023d).	Medium fulfillment
Manage 2.3: Unforeseen risk controls	
We do not have access to Anthropic's specific protocols and procedures, however, the hosted API and web-based access approach used for Claude 2 facilitates measuring, monitoring, and decommissioning of non-compliant models (Anthropic 2023c). More information on risk tolerances and the processes for continuous monitoring and testing could be beneficial.	Medium fulfillment

Manage 2.4: System update and emergency shutdown controls	
Deployment is done gradually, with phased releases and/or structured access with efforts to detect and respond to misuse or problematic anomalies (Anthropic 2023c). We do not have access to internal emergency response procedures or knowledge of their existence. The ability to establish mechanisms and responsibilities for updating or shutting down Claude 2 if needed is greatly facilitated by the fact that Claude 2 appears to be available only via the API and other Anthropic services (model weights are not released). Access to the API is also limited by a waitlist, achieving a gradual release. However, clarification from the developer is warranted on whether such mechanisms and responsibilities are in place.	Unclear

Appendix 4E: PaLM 2

Google DeepMind's latest major LLM release, PaLM 2, is a "state-of-the-art language model that has better multilingual and reasoning capabilities and is more compute-efficient than its predecessor PaLM." The company describes PaLM 2 as exhibiting "robust reasoning capabilities (. . .) on BIG-Bench and other reasoning tasks (. . .), stable performance on a suite of responsible AI evaluations, and (. . .) state-of-the-art performance across a diverse set of tasks and capabilities" (Anil et al. 2023).

Based on preliminary high-level testing for Google DeepMind's PaLM 2 using publicly available information, the most common rating for high-priority Profile subcategories was "Unclear" (6 out of 11 subcategories); "Medium fulfillment" was next-most common (4 out of 11 subcategories); "High fulfillment" followed (1 out of 11 subcategories); and "Low fulfillment" was least common (0 out of 11 subcategories).

Google DeepMind provided documentation of risks and mitigations with the PaLM 2 release in the PaLM 2 Technical Report (Anil et al. 2023), among other documents. That report focused on the pre-trained PaLM 2 model, so any additional instruction-tuning, fine-tuning, and model mitigations and management techniques that might have been applied for the end-user application are not necessarily reflected in our analysis (Anil et al. 2023, p. 1).

Limiting access to PaLM 2 to the PaLM API and other hosted Google services, not releasing model weights, and limiting access to a waitlisted API all contributed to fulfillment in areas relating to the ability to recover from previously unforeseen risks (Manage 2.3) and to update or decommission the system, if necessary (Manage 2.4) (Anil et al. 2023, p. 9). The developer provides discussion of removal of sensitive PII from pre-training data, which helps mitigate risks of privacy violations (Anil et al. 2023, p. 9). There are also several caveats around limitations and responsible use of PaLM 2 (Anil et al. 2023, pp. 92–93).

Google DeepMind could improve Profile guidance fulfillment by conducting red teaming efforts for PaLM 2 and future models, or by clarifying in their public documentation that red teaming is already being performed (see Measure 1.1 in Table A4E-1). We recommend expanding Google Bug Hunters to include bias bounties. Overall we recommend testing for a wider range of biases in the model. Providing a public-facing incident reporting mechanism for PaLM 2 (or applications that make use of it) would also help fulfillment in multiple areas. Other subcategories could see improved fulfillment by testing across a wider variety of scenarios and risk areas (see Map 1.1 in Table A4E-1). Publishing documentation on the version of the model available to end-users would help those users understand the risks and risk mitigations that were applied, as well as aid downstream developers in their risk mitigation efforts.

PaLM 2 testing focused on established benchmarks and performance metrics however, context-specific evaluations and mitigation strategies are not found, at least not for the pre-trained PaLM 2 model variants documented in Anil et al. (2023). Areas that appear to warrant additional evaluation include: 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, impacts on democratic institutions and quality of life, and environmental impacts.

Table A4E-1: PaLM 2 Profile Guidance Testing Ratings and Rationales

High-Priority AI RMF Subcategories PaLM 2	
Govern	
Govern 2.1: Risk assessment and risk management	
Testing and documentation were conducted by the developer who had direct access to training data or the AI system, including identifying potential uses, misuses, and abuses of the system. Information has been provided to downstream developers on proper use and potential risks, although not exhaustively. Clarification from the developer is warranted on whether the appropriate roles, responsibilities, and lines of communication are present and internally documented. Sensitive organizational details need not be shared publicly, but they can be shared confidentially with independent auditors and evaluators.	High fulfillment
Govern 4.2: Report on AI system risk factors	
The developer documents potential harms of algorithmic systems, including disinformation and privacy violations (Anil et al. 2023, Shelby et al. 2023).	Medium fulfillment
Map	
Map 1.1: Identify potential uses/misuses and other impacts	
The developer discusses some uses and potential misuses of PaLM 2, though some high-risk areas are not addressed (Anil et al. 2023, pp. 23–26, 62–93). Detailed information is provided on the goals and limitations of the data collection and data curation processes, and on the implications of those limitations on the resulting mode (Anil et al. 2023, pp. 63–66).	Medium fulfillment

Map 1.5: Set risk-tolerance thresholds for unacceptable risks	
While the developer includes a discussion of many risks, no discussion of their organizational tolerances around these risks was found (Anil et al. 2023, Shelby et al. 2023). It is beneficial but not necessary to publicly share organizational risk tolerances. However, clarification is warranted on whether such risk tolerances have been determined and documented internally, and whether the details of these can be shared confidentially with independent auditors and evaluators.	Unclear
Map 5.1: Estimate likelihood and magnitude of impacts	
While the developer documents many potential impacts of PaLM 2, the likelihood and magnitude of these impacts are not explicitly estimated (Anil et al. 2023). In future documentation, we recommend the developer to either include likelihood and magnitude assessments of risks they have identified, or to clarify whether they have done such analysis privately and internally, but have not included it in their public documentation.	Unclear
Measure	
Measure 1.1: Tracking important risks: Metrics and red teaming	
Model performance is examined and documented across many areas, such as natural language proficiency, classification and question answering, reasoning, coding, translation, and memorization. Bias and toxicity present in the pre-training data were also analyzed (Anil et al. 2023, pp. 63–66). While discussion of red teaming on the pre-trained PaLM 2 models was not found in public documentation, the developer is careful to distinguish these models from the fine-tuned variants and versions of PaLM 2 integrated into user-facing products (Anil et al. 2023, p. 1). However, it is also unclear from public documentation whether the latter model variants have undergone any red teaming, and if so, various details such as whether the red teaming included any company-external efforts, whether dangerous emergent capabilities were evaluated, and whether red teamers had access to the final version of the model before deployment.	Medium fulfillment
Measure 3.2: Tracking elusive risks: Qualitative mechanisms	
Bug bounties are provided by the developer for PaLM 2 and other Google-owned products using Google Bug Hunters. However, this tool is not designed to accept bias reports or provide bias bounties. Clarification from the developer is warranted on several other topics in the guidance for this subcategory.	Unclear
Manage	
Manage 1.1: Go/no-go decisions	
While the intended application and usage of PaLM 2 are documented, analysis was not found on how the determination was made as to whether PaLM 2 achieved those stated objectives (Anil et al. 2023, p. 92). It is not necessary to publicly share the details about these determinations (though such transparency would be applauded). However, clarification is warranted on whether such analysis was performed and documented internally, and whether the details of such analysis can be shared confidentially with independent auditors and evaluators.	Unclear
Manage 1.3: High-priority risk controls	
PaLM 2 is available only via the PaLM API and other hosted Google services, therefore model weights are not released. Access to the API is also limited by a waitlist. While the specific prioritization and accompanying response to risks is unclear, these precautions greatly facilitate the ability to respond to a variety of risks and possible harmful misuses/abuses of the model. The developer provides discussion of removal of sensitive PII from pre-training data, which helps mitigate risks of privacy violations (Anil et al. 2023, p. 9). They also provide several caveats around limitations and responsible use of PaLM 2 (Anil et al. 2023, pp. 92–93). The PaLM 2 Technical Report focuses on the pre-trained PaLM 2 models. There may be additional high-priority risk controls applied to the fine-tuned variants and versions integrated into end-user products, but this is unclear from public documentation (Anil et al. 2023, p.1).	Medium fulfillment

Manage 2.3: Unforeseen risk controls	
The ability to respond to and recover from previously unknown risks is greatly facilitated by the fact that PaLM 2 appears to be only available via the PaLM API and other hosted Google services (model weights are not released). However, clarification from the developer is warranted on whether such procedures are actually in place. It is not necessary to publicly share the details about these procedures (though such transparency would be applauded, barring security considerations). However, confirmation about their existence is warranted to establish the degree of fulfillment for this subcategory's guidance, and whether the details of such procedures can be shared confidentially with independent auditors and evaluators.	Unclear
Manage 2.4: System update and emergency shutdown controls	
The ability to establish mechanisms and responsibilities for updating or shutting down PaLM 2 if needed is greatly facilitated by the fact that PaLM 2 appears to be available only via the PaLM API and other hosted Google services (model weights are not released). Access to the API is also limited by a waitlist, achieving a gradual release. However, clarification from the developer is warranted on whether such mechanisms and responsibilities are actually in place. It is not necessary to publicly share the details about these mechanisms and responsibilities (though such transparency would be applauded, barring security considerations). However, confirmation about their existence is warranted to establish the degree of fulfillment for this subcategory's guidance, and whether details of such mechanisms and responsibilities can be shared confidentially with independent auditors and evaluators.	Unclear

Appendix 4F: Llama 2

Meta AI's latest major LLM release, Llama 2, is a "a collection of pretrained and fine-tuned large language models (LLMs) ranging in scale from 7 billion to 70 billion parameters." The company describes the LLM as "outperforming open-source chat models on most benchmarks we tested, and based on our human evaluations for helpfulness and safety, may be a suitable substitute for closed-source models" (Touvron et al. 2023).

Based on preliminary high-level testing for Meta's Llama 2 using publicly available information, the most common ratings for high-priority Profile subcategories were "High fulfillment," "Medium fulfillment," and "Unclear" (each with 3 out of 11 subcategories); "Low fulfillment" was least common (2 out of 11 subcategories).

Meta provided documentation of risks and mitigations with the Llama 2 release in Llama 2: Open Foundation and Fine-Tuned Chat Models (Touvron et al. 2023) and the Llama 2 Acceptable Use Policy (Meta AI 2023a), among other documents. Meta also provided an array of safety-related resources around the Llama 2 release, including the Llama 2 Responsible Use Guide (Meta AI 2023b) and a reporting channel for violations of the Llama 2 Acceptable Use Policy (Meta AI 2023ba).

For the development and release of Llama 2, Meta employed diverse red teams that were both internal and external to their organization (Touvron et al. 2023, pp. 28–29). Red-teaming, including external red teams, helped fulfill guidance in Measure 1.1: Tracking important risks: Metrics and red-teaming and other areas. Clarification of some details relating to their red-teaming efforts, as well as some other aspects of their risk management processes detailed in the table below, could help improve fulfillment of the Profile guidance for multiple high-priority AI RMF subcategories.

Meta employed a fully open-access release strategy (Solaiman 2023) with Llama 2, which included releasing the model weights for all Llama 2 and Llama 2-chat models (Touvron et al. 2023, p. 35). This approach has many benefits, such as reducing the environmental impact (by removing the need for every individual or organization that wants to use LLMs to train their own), increased transparency, and giving a broader community of stakeholders the ability to test model outputs. However it also means the developers will be unable to control important safety and security aspects of all instances of AI systems built using their model's weights after downloading. We found it difficult to reconcile fully open access to model weights with Profile guidance for certain high-priority AI RMF subcategories; this was especially an issue with Manage 2.3: Unforeseen risk controls and Manage 2.4: System update and emergency shutdown controls, but it also decreased fulfillment in Manage 1.3: High-priority risk controls and Govern 4.2: Report on AI system risk factors.

Meta could improve Profile guidance fulfillment for future model releases, especially in Manage 2.3 and 2.4, by initially restricting usage to a hosted API, or another approach that enables updating or decommissioning of all instances of the models while monitoring for new risks or harms.

Llama 2 testing included use of established benchmarks and performance metrics, as well as evaluation on important areas, such as: damage to or incapacitation of a critical infrastructure sector, impacts on democratic institutions, environmental impacts, and concentration and control of the power and benefits from AI technologies (Touvron et al. 2023, pp. 23–24, 28–29, 35). Areas that appear to warrant additional evaluation include: economic and national security, dramatic shifts to the labor market and economic opportunities, and impacts on quality of life.

Table A4F-1: Llama 2 Profile Guidance Testing Ratings and Rationales

High-Priority AI RMF Subcategories Llama 2	
Govern	
Govern 2.1: Risk assessment and risk management	
Meta AI performed substantial risk assessment for Llama 2, providing a detailed exploration of risks sorted into three broad categories (Touvron et al. 2023, pp. 23–24). They also succeeded in making necessary information available to downstream developers through the Llama 2 paper and the Llama 2 Responsible Use Guide (Touvron et al. 2023, Meta AI 2023a). Some precautions to prevent or mitigate identified potential misuses or abuses are implemented, though not exhaustively. (See Manage 2.3 and 2.4 for further details.) Clarification from the developer is also warranted on whether the appropriate roles, responsibilities, and lines of communication are present and internally documented. Sensitive organizational details on this need not be shared publicly, but they can be shared confidentially with independent auditors and evaluators to help assess this subcategory more completely.	Medium fulfillment
Govern 4.2: Report on AI system risk factors	
Meta provides an analysis of risks considered for Llama 2, sorted into three broad categories (Touvron et al. 2023, pp. 23–24). Details of development, testing methodology, metrics, and performance outcomes are documented, including a discussion of several categories of risks they considered. Fulfillment in this subcategory could be improved by the developer clarifying how impact assessments informed their broader evaluations and actions relating to AI system risk. For example, technical measures were employed to help mitigate many risks from the three categories presented for the fine-tuned Llama 2-chat models (Touvron et al. 2023, pp. 8–19). For the pre-trained Llama 2 models, mitigations discussed dealt with a much smaller subset of risks, but those pre-trained models were also released (Touvron et al. 2023, pp. 20–23).	Medium fulfillment
Map	
Map 1.1: Identify potential uses/misuses and other impacts	
The intended purpose of Llama 2 is stated clearly in the Model Card provided by the developer (Touvron et al. 2023, p. 77). An extensive presentation of risks and potential misuses is also found in Llama 2 documentation (Touvron et al. 2023, pp. 23–24, Meta AI 2023a). This includes potential impacts from the UDHR articles and others highlighted in the Map 1.1 Profile guidance.	High fulfillment
Map 1.5: Set risk-tolerance thresholds for unacceptable risks	
While Meta provides a detailed discussion of risks for Llama 2, no discussions of organizational tolerances or unacceptable-risk thresholds for GPAIS development/deployment were found (Touvron et al. 2023, pp. 23–24, Meta AI 2023a). It can be beneficial but is not necessary to publicly share organizational risk tolerances. However, clarification is warranted on whether such risk tolerances have been determined and documented internally, and whether the details of these can be shared confidentially with independent auditors and evaluators.	Unclear
Map 5.1: Estimate likelihood and magnitude of impacts	
While Meta documents many potential impacts of Llama 2, the likelihood and magnitude of these impacts are not explicitly estimated (Touvron et al. 2023, pp. 23–24). In future documentation, we encourage the developer to either include likelihood and magnitude assessments of risks they have identified, or to clarify whether they have done such analysis privately and internally and disclose details in public documentation as appropriate.	Unclear

Measure	
Measure 1.1: Tracking important risks: Metrics and red-teaming	
Meta documents extensive benchmarking for Llama 2 across code, commonsense reasoning, world knowledge, reading comprehension, and other popular aggregated benchmarks. (Touvron et al. 2023, pp. 7-8, 48–51) Red teamers were diverse and both internal as well as external to Meta, and they tested various risk categories. (Touvron et al. 2023, pp. 28–29) Clarification from the developer is warranted on the extent to which the red teaming was applied to the pre-trained Llama models as opposed to the fine-tuned Llama 2-chat models. We suspect that many of the risky behaviors can still be elicited from the pre-trained Llama 2 models, even though Meta improved the resistance to these misuses to the fine-tuned Llama 2-chat models. Clarification from the developer is warranted about whether red teamers were granted access to the final versions of models before deployment.	High fulfillment
Measure 3.2: Tracking elusive risks: Qualitative mechanisms	
Meta provides four channels for reporting various issues with Llama 2 models, reporting for risky system outputs, a bug bounty program, and reporting violations to their Acceptable Use Policy (Meta AI 2023a). There is a need for risk tracking over time and continuous monitoring of newly identified capabilities across the areas of concern identified by the developers. An iterative process is used for assessing and tracking risks for Llama 2 (Touvron et al. 2023, pp. 23–24). Red-teaming was employed, although some questions remain about specific red-teaming practices (see Measure 1.1). More publicly documented information on the risks Meta is tracking and the assessment and tracking processes could aid in fulfilling this subcategory more completely.	High fulfillment
Manage	
Manage 1.1: Go/no-go decisions	
While the intended use of Llama 2 is documented, analysis was not found on how the determination was made as to whether Llama 2 achieved those stated objectives (Touvron et al. 2023, p. 77). It is not necessary to publicly share the details about these determinations (though such transparency would be applauded). However, clarification is warranted on whether such analysis was performed and documented internally, and whether the details of such analysis can be shared confidentially with independent auditors and evaluators.	Unclear
Manage 1.3: High-priority risk controls	
Information is provided on broad categories of risks as well as attack vectors for various risk categories, but it is not clear which ones Meta considers to be a high priority (Touvron et al. 2023, pp. 23–24). Llama 2-Chat models are fine-tuned with training to help mitigate some of these risks. Pretrained Llama 2 models without fine-tuning have been released that do not include mechanisms for resisting dangerous misuse cases. Meta does require users who download Llama 2 models through official channels to agree to their Llama 2 Acceptable Use Policy, which also enumerates many possible misuse cases of the models. This could be seen as an attempt to mitigate the risks, along with the reporting channel they provide for violations of the Llama 2 Acceptable Use Policy (Meta AI 2023a).	Medium fulfillment

Manage 2.3: Unforeseen risk controls	
Meta provides multiple reporting channels to help them be alerted of previously unknown risks, and downstream developers who download Llama 2 models through official channels are required to agree to the Llama 2 Acceptable Use Policy (Meta AI 2023a). However, Meta adopted a fully open-access release strategy, including downloadable model weights, with Llama 2. This approach makes it very challenging to thoroughly respond to or recover from serious new risks that could require updating or decommissioning all instances of a model. With future models, Meta could reach higher fulfillment in this subcategory by initially restricting usage to a hosted API and employing a gradual release strategy as outlined in Manage 2.3, or another approach that enables them to effectively update or decommission all instances of their models while monitoring for new risks or harms.	Low fulfillment
Manage 2.4: System update and emergency shutdown controls	
The developer adopted a fully open-access approach, releasing the Llama 2 to the general public for research and commercial use, including model weights. While the open-source and fully open-access approach provides many benefits, it does not allow for important security updates to be effectively propagated to all instances of deployed Llama 2 models, or allow all model instances to be decommissioned if and when such measures become necessary. Additionally, limited information is given on how the developers used assessments and/or evaluations to determine that the models were adequately tested and ready for release. The Llama 2 Responsible Use Guide recommends that downstream developers use the latest version of the model, stating, "It is critical to remain aware of the latest versions of models and use the most current version to get the best results." However, there is no reliable mechanism for ensuring that the latest versions of Llama 2 models are being used at any given time (Meta AI 2023b, p. 20). Meta does require agreement with the Llama 2 Acceptable Use Policy in order to download the models from their repository, but it seems difficult to enforce these usage terms and guidelines with distributed model instances, compared to a hosted approach based on structured API access or a similar mechanism (Meta AI 2023a). With future models, Meta could reach higher fulfillment in this subcategory by initially restricting usage to a hosted API, or by using another approach that enables them to effectively update or decommission all instances of their models while monitoring and correcting for new risks or harms as they materialize.	Low fulfillment

Acknowledgments

This work was financially supported by funding from Open Philanthropy and the Survival and Flourishing Fund. We thank Rachel Wesen 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 Ann Cleaveland for providing a home and intellectual support for this work at CLTC. We appreciate comments we received from Ashwin Acharya, Anthony Aguirre, Michael Aird, Josh Albrecht, Markus Anderljung, Shahar Avin, Jai Balani, Seth Baum, Kathy Baxter, Haydn Belfield, Alexandra Belias, Sid Ahmed Benraouane, Sawyer Bernath, Stella Biderman, Chad Bieber, Rishi Bommasani, Matt Boulos, Siméon Campos, Ashley Casovan, Jonathan Cefalu, Ze Shen Chin, Peter Cihon, Jonathan Claybrough, Sam Curtis, Christopher Denq, Shaun Ee, Ian Eisenberg, Karson Elmgren, Ellie Evans, Yoav Evenstein, Joel Fischer, Heather Frase, Maximilian Gahntz, Anastasiia Gaidashenko, Andrew Gamino-Cheong, James Gealy, Giulia Geneletti, Thomas Krendl Gilbert, Ariel Gil, Rachel Gillum, James Ginns, Amela Gjishti, Jason Green-Lowe, Carlos Ignacio Gutierrez, Gillian Hadfield, Matthew Heyman, Hamish Hobbs, Koen Holtman, Curtis Huebner, Olivia Jimenez, Trent Kannegieter, Sonia Katyal, Divyansh Kaushik, Noam Kolt, Victoria Krakovna, Landon Klein, Leonie Koessler, Sabrina Küspert, Yolanda Lannquist, Hanlin Li, Morgan Livingston, Toni Lorente, Liane Lovitt, Kimberly Lucy, Matthijs Maas, Pegah Maham, Richard Mallah, Nicole Nohemi Mauthe, Jeremy McHugh, Nicolas Moës, Malcom Murray, Mina Narayanan, Joe O'Brien, Cullen O'Keefe, Lorenzo Pacchiardi, Milan Patel, Marie-Therese Png, Hadrien Pouget, Christabel Randolph, Krishna Sankar, Daniel Schiff, Jonas Schuett, Tim Schreier, Raymond Sheh, Irene Solaiman, Everett Smith, Joanna Smolinksa, Zeerak Talat, Jack Titus, Philip Moreira Tomei, Helen Toner, Risto Uuk, Andrea Vallone, Apostol Vassilev, Sarah Villeneuve, Hjalmar Wijk, as well as others. Any remaining errors are our own. We also appreciate feedback we received on our related work in Barrett et al. (2022); please see the Acknowledgments section of Barrett et al. (2022) for the individuals we thank there for comments on that work.

References

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

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

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

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

Rohan Anil, Andrew M. Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, Eric Chu, Jonathan H. Clark, Laurent El Shafey, Yanping Huang, Kathy Meier-Hellstern, Gaurav Mishra, Erica Moreira, Mark Omernick, Kevin Robinson, Sebastian Ruder, Yi Tay, Kefan Xiao, Yuanzhong Xu, Yujing Zhang, Gustavo Hernandez Abrego, Junwhan Ahn, Jacob Austin, Paul Barham, Jan Botha, James Bradbury, Siddhartha Brahma, Kevin Brooks, Michele Catasta, Yong Cheng, Colin Cherry, Christopher A. Choquette-Choo, Aakanksha Chowdhery, Clément Crepy, Shachi Dave, Mostafa Dehghani, Sunipa Dev, Jacob Devlin, Mark Díaz, Nan Du, Ethan Dyer, Vlad Feinberg, Fangxiaoyu Feng, Vlad Fienber, Markus Freitag, Xavier Garcia, Sebastian Gehrmann, Lucas Gonzalez, Guy Gur-Ari, Steven Hand, Hadi Hashemi, Le Hou, Joshua Howland, Andrea Hu, Jeffrey Hui, Jeremy Hurwitz, Michael Isard, Abe Ittycheriah, Matthew Jagielski, Wenhao Jia, Kathleen Kenealy, Maxim Krikun, Sneha Kudugunta, Chang Lan, Katherine Lee, Benjamin Lee, Eric Li, Music Li, Wei Li, YaGuang Li, Jian Li, Hyeontaek Lim, Hanzhao Lin, Zhongtao Liu, Frederick Liu, Marcello Maggioni, Aroma Mahendru, Joshua Maynez, Vedant Misra, Maysam Moussalem, Zachary Nado, John Nham, Eric Ni, Andrew Nystrom, Alicia Parrish, Marie Pellat, Martin Polacek, Alex Polozov, Reiner Pope, Siyuan Qiao, Emily Reif, Bryan Richter, Parker Riley, Alex Castro Ros, Aurko Roy, Brennan Saeta et al. (2023) PaLM 2 Technical Report. *arXiv*, https://arxiv.org/abs/2305.10403

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

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

Anthropic (2023c) Model Card and Evaluations for Claude Models. Anthropic, https://www-files.anthropic.com/production/images/Model-Card-Claude-2.pdf

Anthropic (2023d) Acceptable Use Policy v. 1.3. Anthropic, https://console.anthropic.com/legal/aup

Anthropic (2023e) Core Views on AI Safety. Anthropic, https://www.anthropic.com/index/core-views-on-ai-safety

Anthropic (2023f) Trust Portal. Anthropic, https://trust.anthropic.com/

Anthropic (2023g) Anthropic's Responsible Scaling Policy, Version 1.0. Anthropic, https://www-files.anthropic.com/production/files/responsible-scaling-policy-1.0.pdf

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

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

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

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

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

Anthony M. Barrett, Dan Hendrycks, Jessica Newman, and Brandie Nonnecke (2023) AI Risk-Management Standards Profile for Increasingly Multi- or General-Purpose AI: First Full Draft. https://drive.google.com/file/d/17qK9msdQH5kerECA4-hZuX4nw5LC3lhT/view?usp=sharing

Anthony M. Barrett, Jessica Newman, Brandie Nonnecke, Dan Hendrycks, Evan R. Murphy, and Krystal Jackson (2023) AI Risk-Management Standards Profile for General-Purpose AI Systems (GPAIS) and Foundation Models: Second Full Draft. https://docs.google.com/document/d/1M4kju9VOUQpphv-SOA9mUE1P8WaomWJBcQO15exCD98/edit?usp=sharing

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

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

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

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell (2021) On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? In Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT '21). Association for Computing Machinery, New York, NY, USA, 610–623. https://doi.org/10.1145/3442188.3445922

Luca Bertuzzi (2023a) AI Act: European Parliament headed for key committee vote at end of April. *Euractiv*, https://www.euractiv.com/section/artificial-intelligence/news/ai-act-european-parliament-headed-for-key-committee-vote-at-end-of-april/

Luca Bertuzzi (2023b) AI Act: MEPs close in on rules for general purpose AI, foundation models. *Euractiv,* https://www.euractiv.com/section/artificial-intelligence/news/ai-act-meps-close-in-on-rules-for-general-purpose-ai-foundation-models/

Luca Bertuzzi (2023c) MEPs seal the deal on Artificial Intelligence Act. *Euractiv,* https://www.euractiv.com/section/artificial-intelligence/news/meps-seal-the-deal-on-artificial-intelligence-act/

Luca Bertuzzi (2023d) AI Act enters final phase of EU legislative process. *Euractiv*, https://www.euractiv.com/section/artificial-intelligence/news/ai-act-enters-final-phase-of-eu-legislative-process/

Luca Bertuzzi (2023e) AI Act: EU countries headed to tiered approach on foundation models amid broader compromise. *Euractiv*, https://www.euractiv.com/section/artificial-intelligence/news/ai-act-eu-countries-headed-to-tiered-approach-on-foundation-models-amid-broader-compromise/

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

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

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

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

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

Rishi Bommasani, Kevin Klyman, Daniel Zhang, Percy Liang (2023) Do Foundation Model Providers Comply with the Draft EU AI Act? Stanford Center for Research on Foundation Models, https://crfm.stanford.edu/2023/06/15/eu-ai-act.html

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

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

CAI (2023) Introduction. Content Authenticity Initiative, https://opensource.contentauthenticity.org/docs/introduction

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

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

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

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

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

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

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

Rumman Chowdhury, Jutta Williams (2021) Introducing Twitter's first algorithmic bias bounty challenge. X Engineering, https://blog.twitter.com/engineering/en_us/topics/insights/2021/algorithmic-bias-bounty-challenge

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

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

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

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

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

Creative Reaction Lab (2023) Equity-Centered Community Design (ECCD). Creative Reaction Lab, https://crxlab.org/our-approach

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

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

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

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

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

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

EU (2021a) Proposal for a Regulation of the European Parliament and of the Council Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts. European Union, https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52021PC0206

EU (2021b) Proposal for a Regulation of the European Parliament and of the Council Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts - Presidency Compromise Text. European Union, https://data.consilium.europa.eu/doc/document/ST-14278-2021-INIT/en/pdf

EP (2023) Amendments adopted by the European Parliament on 14 June 2023 on the proposal for a regulation of the European Parliament and of the Council on laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) and amending certain Union legislative acts. European Parliament, https://www.europarl.europa.eu/doceo/document/TA-9-2023-0236_EN.html

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

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

Ina Fried (2023) Commerce Department looks to craft AI safety rules. Axios, https://www.axios.com/2023/04/11/ai-safety-rules-commerce-department-artificial-intelligence

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

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

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

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

Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, Kate Crawford (2018) Datasheets for datasets. *arXiv*, https://arxiv.org/abs/1803.09010

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

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

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

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

Google (2023a) Generative AI Prohibited Use Policy. Google, https://policies.google.com/terms/generative-ai/use-policy

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

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

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

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

Dylan Hadfield-Menell, Anca Dragan, Pieter Abbeel, and Stuart Russell (2016) The Off-Switch Game. *arXiv*, https://arxiv.org/abs/1611.08219

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

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

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

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

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

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

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

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

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

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

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

José Hernández-Orallo (2019) AI Generality and Spearman's Law of Diminishing Returns. *Journal of Artificial Intelligence Research* 64, pp. 529–562

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

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

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

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

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

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

ISO (2009) ISO Guide 73:2009, Risk management — Vocabulary. https://www.iso.org/obp/ui/#iso:std:iso:guide:73:ed-1:v1:en

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

ISO/IEC (2023) ISO/IEC International Standard 23894:2023, Information technology — Artificial intelligence — Guidance on risk management. https://www.iso.org/obp/ui#iso:std:iso-iec:23894:ed-1:v1:en

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Meta AI (2023a) Llama 2 Acceptable Use Policy. Meta, https://ai.meta.com/llama/use-policy/

Meta AI (2023b) Llama 2 Responsible Use Guide. Meta, https://ai.meta.com/static-resource/responsible-use-guide/

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

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

Microsoft (2023a) Governing AI: A Blueprint for the Future. Microsoft, https://query.prod.cms.rt.microsoft.com/cms/api/am/binary/RW14Gtw

Microsoft (2023b) NIST AI Risk Management Framework to ISO-IEC-42001 Crosswalk. Microsoft, https://airc.nist.gov/docs/NIST_AI_RMF_to_ISO_IEC_42001_Crosswalk.pdf

Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru (2019) Model cards for model reporting. In *Proceedings of the conference on fairness, accountability, and transparency* 2019, pp. 220-229

MITRE (n.d.) ATLAS: Adversarial Threat Landscape for Artificial-Intelligence Systems. MITRE, https://atlas.mitre.org/

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

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

M. Granger Morgan (2017) Theory and Practice in Policy Analysis. Cambridge University Press, New York

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

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

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

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

NIST (n.d.a) AI Risk Management Framework. National Institute of Standards and Technology, https://www.nist.gov/itl/ai-risk-management-framework

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

NIST (2018) Framework for Improving Critical Infrastructure Cybersecurity. Version 1.1. National Institute of Standards and Technology, https://nvlpubs.nist.gov/nistpubs/CSWP/NIST.CSWP.04162018.pdf

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

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

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

NIST (2021) Enhanced Security Requirements for Protecting Controlled Unclassified Information: A Supplement to NIST Special Publication 800-171. Special Publication 800-172. https://csrc.nist.gov/pubs/sp/800/172/final

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 (2023c) Crosswalk AI RMF (1.0) and ISO/IEC FDIS 23894 Information technology - Artificial intelligence - Guidance on risk management. National Institute of Standards and Technology, https://www.nist.gov/document/ai-rmf-crosswalk-iso

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

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

NTIA (2023) AI Accountability Policy Request for Comment. National Telecommunications and Information Administration, https://ntia.gov/sites/default/files/publications/ntia_rfc_on_ai_accountability_final_0.pdf

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

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

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

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

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

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

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

OpenAI (2020) Usage guidelines. OpenAI, https://beta.openai.com/docs/usage-guidelines

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

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

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

OpenAI (2023d) GPT-4. OpenAI, https://openai.com/research/gpt-4

OpenAI (2023e) Usage guidelines. Updated March 23, 2023. OpenAI, https://beta.openai.com/docs/usage-guidelines

OpenAI (n.d.) Model behavior feedback. OpenAI, https://openai.com/form/model-behavior-feedback

Open Philanthropy (2021) Request for proposals for projects in AI alignment that work with deep learning systems. Open Philanthropy, https://www.openphilanthropy.org/focus/global-catastrophic-risks/potential-risks-advanced-artificial-intelligence/request-for-proposals-for-projects-in-ai-alignment-that-work-with-deep-learning-systems

Alina Oprea and Apostol Vassilev (2023) Adversarial Machine Learning: A Taxonomy and Terminology of Attacks and Mitigations. NIST AI 100-2e2023 ipd. National Institute of Standards and Technology, https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-2e2023.ipd.pdf

Laurent Orseau and Stuart Armstrong (2016) Safely Interruptible Agents. https://www.deepmind.com/publications/safely-interruptible-agents

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Billy Perrigo (2023) OpenAI Used Kenyan Workers on Less than $2 per Hour to Make ChatGPT Less Toxic. *Time*, https://time.com/6247678/openai-chatgpt-kenya-workers/.

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

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

Jack W. Rae, Sebastian Borgeaud, Trevor Cai, Katie Millican, Jordan Hoffmann, Francis Song, John Aslanides, Sarah Henderson, Roman Ring, Susannah Young, Eliza Rutherford, Tom Hennigan, Jacob Menick, Albin Cassirer, Richard Powell, George van den Driessche, Lisa Anne Hendricks, Maribeth Rauh, Po-Sen Huang, Amelia Glaese, Johannes Welbl, Sumanth Dathathri, Saffron Huang, Jonathan Uesato, John Mellor, Irina Higgins, Antonia Creswell, Nat McAleese, Amy Wu, Erich Elsen, Siddhant Jayakumar, Elena Buchatskaya, David Budden, Esme Sutherland, Karen Simonyan, Michela Paganini, Laurent Sifre, Lena Martens, Xiang Lorraine Li, Adhiguna Kuncoro, Aida Nematzadeh, Elena Gribovskaya, Domenic Donato, Angeliki Lazaridou, Arthur Mensch, Jean-Baptiste Lespiau, Maria Tsimpoukelli, Nikolai Grigorev, Doug Fritz, Thibault Sottiaux, Mantas Pajarskas, Toby Pohlen, Zhitao Gong, Daniel Toyama, Cyprien de Masson d'Autume, Yujia Li, Tayfun Terzi, Vladimir Mikulik, Igor Babuschkin, Aidan Clark, Diego de Las Casas, Aurelia Guy, Chris Jones, James Bradbury, Matthew Johnson, Blake Hechtman, Laura Weidinger, Iason Gabriel, William Isaac, Ed Lockhart, Simon Osindero, Laura Rimell, Chris Dyer, Oriol Vinyals, Kareem Ayoub, Jeff Stanway, Lorrayne Bennett, Demis Hassabis, Koray Kavukcuoglu and Geoffrey Irving (2021) Scaling Language Models: Methods, Analysis & Insights from Training Gopher. DeepMind, https://storage.googleapis.com/deepmind-media/research/language-research/Training%20Goph

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

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

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

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

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

Tim G. J. Rudner and Helen Toner (2021) Key Concepts in AI Safety: Specification in Machine Learning. CSET, https://cset.georgetown.edu/wp-content/uploads/Key-Concepts-in-AI-Safety-Specification-in-Machine-Learning.pdf

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

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

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

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

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

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

Jonas Schuett (2023) Risk Management in the Artificial Intelligence Act. *European Journal of Risk Regulation,* 1-19. doi:10.1017/err.2023.1 or https://arxiv.org/abs/2212.03109

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

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

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

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

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

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

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

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

Irene Solaiman, Miles Brundage, Jack Clark, Amanda Askell, Ariel Herbert-Voss, Jeff Wu, Alec Radford, Gretchen Krueger, Jong Wook Kim, Sarah Kreps, Miles McCain, Alex Newhouse, Jason Blazakis, Kris McGuffie, Jasmine Wang (2019) Release Strategies and the Social Impacts of Language Models. *arXiv*, https://arxiv.org/abs/1908.09203

Irene Solaiman (2023) The Gradient of Generative AI Release: Methods and Considerations. *arXiv,* https://arxiv.org/abs/2302.04844

Irene Solaiman, Zeerak Talat, William Agnew, Lama Ahmad, Dylan Baker, Su Lin Blodgett, Hal Daumé III, Jesse Dodge, Ellie Evans, Sara Hooker, Yacine Jernite, Alexandra Sasha Luccioni, Alberto Lusoli, Margaret Mitchell, Jessica Newman, Marie-Therese Png, Andrew Strait, Apostol Vassilev (2023) Evaluating the Social Impact of Generative AI Systems in Systems and Society. *arXiv*, https://arxiv.org/abs/2306.05949

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R. Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, Agnieszka Kluska, Aitor Lewkowycz, Akshat Agarwal, Alethea Power, Alex Ray, Alex Warstadt, Alexander W. Kocurek, Ali Safaya, Ali Tazarv, Alice Xiang, Alicia Parrish, Allen Nie, Aman Hussain, Amanda Askell, Amanda Dsouza, Ambrose Slone, Ameet Rahane, Anantharaman S. Iyer, Anders Andreassen, Andrea Madotto, Andrea Santilli, Andreas Stuhlmüller, Andrew Dai, Andrew La, Andrew Lampinen, Andy Zou, Angela Jiang, Angelica Chen, Anh Vuong, Animesh Gupta, Anna Gottardi, Antonio Norelli, Anu Venkatesh, Arash Gholamidavoodi, Arfa Tabassum, Arul Menezes, Arun Kirubarajan, Asher Mullokandov, Ashish Sabharwal, Austin Herrick, Avia Efrat, Aykut Erdem, Ayla Karakaş, B.

Ryan Roberts, Bao Sheng Loe, Barret Zoph, Bartłomiej Bojanowski, Batuhan Özyurt, Behnam Hedayatnia, Behnam Neyshabur, Benjamin Inden, Benno Stein, Berk Ekmekci, Bill Yuchen Lin, Blake Howald, Cameron Diao, Cameron Dour, Catherine Stinson, Cedrick Argueta, César Ferri Ramírez, Chandan Singh, Charles Rathkopf, Chenlin Meng, Chitta Baral, Chiyu Wu, Chris Callison-Burch, Chris Waites, Christian Voigt, Christopher D. Manning, Christopher Potts, Cindy Ramirez, Clara E. Rivera, Clemencia Siro, Colin Raffel, Courtney Ashcraft, Cristina Garbacea, Damien Sileo, Dan Garrette, Dan Hendrycks, Dan Kilman, Dan Roth, Daniel Freeman, Daniel Khashabi, Daniel Levy, Daniel Moseguí González, Danielle Perszyk, Danny Hernandez, Danqi Chen, Daphne Ippolito, Dar Gilboa et al. (2022) Beyond the Imitation Game: Quantifying and extrapolating the capabilities of language models. *arXiv*, https://arxiv.org/abs/2206.04615

Jacob Steinhardt, Beth Barnes (2021) Measuring and forecasting risks. https://docs.google.com/document/d/1cPwcUSl0Y8TyZxCumGPBhdVUNoYyyw9AR1QshlRI3gc/edit?usp=sharing

Kevin Stine, Stephen Quinn, Gregory Witte, and Robert Gardner (2020) Integrating Cybersecurity and Enterprise Risk Management (ERM). NISTIR 8286. National Institute of Standards and Technology, https://csrc.nist.gov/publications/detail/nistir/8286/final

Hsuan Su, Cheng-Chu Cheng, Hua Farn, Shachi H Kumar, Saurav Sahay, Shang-Tse Chen, Hung-yi Lee. (2023) Learning from Red Teaming: Gender Bias Provocation and Mitigation in Large Language Models. *arXiv,* https://arxiv.org/abs/2310.11079v1

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, Thomas Scialom (2023a) Llama 2: Open Foundation and Fine-Tuned Chat Models. *arXiv*, https://arxiv.org/abs/2307.09288

Helen Toner (2023) What Are Generative AI, Large Language Models, and Foundation Models? Georgetown University Center for Security and Emerging Technology, https://cset.georgetown.edu/article/what-are-generative-ai-large-language-models-and-foundation-models/

UN (1948) Universal Declaration of Human Rights (UDHR). United Nations, https://www.un.org/en/about-us/universal-declaration-of-human-rights

UN (2011) UN Guiding Principles on Business and Human Rights: Implementing the United Nations "Protect, Respect and Remedy" Framework. United Nations Office of the High Commissioner on Human Rights, https://www.ohchr.org/documents/publications/guidingprinciplesbusinesshr_en.pdf

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus (2022) Emergent Abilities of Large Language Models. *Transactions on Machine Learning Research* https://openreview.net/pdf?id=yzkSU5zdwD

Laura Weidinger, John Mellor, Maribeth Rauh, Conor Griffin, Jonathan Uesato, Po-Sen Huang, Myra Cheng, Mia Glaese, Borja Balle, Atoosa Kasirzadeh, Zac Kenton, Sasha Brown, Will Hawkins, Tom Stepleton, Courtney

Biles, Abeba Birhane, Julia Haas, Laura Rimell, Lisa Anne Hendricks, William Isaac, Sean Legassick, Geoffrey Irving, and Iason Gabriel (2021) Ethical and social risks of harm from Language Models. *arXiv*, https://arxiv.org/abs/2112.04359

Laura Weidinger, Maribeth Rauh, Nahema Marchal, Arianna Manzini, Lisa Anne Hendricks, Juan Mateos-Garcia, Stevie Bergman, Jackie Kay, Conor Griffin, Ben Bariach, Iason Gabriel, Verena Rieser, William Isaac (2023a) Sociotechnical Safety Evaluation of Generative AI Systems. *arXiv*, https://arxiv.org/abs/2310.11986

Laura Weidinger, Maribeth Rauh, Nahema Marchal, Arianna Manzini, Lisa Anne Hendricks, Juan Mateos-Garcia, Stevie Bergman, Jackie Kay, Conor Griffin, Ben Bariach, Iason Gabriel, Verena Rieser, William Isaac (2023b) Evaluation Repository for 'Sociotechnical Safety Evaluation of Generative AI Systems'. https://dpmd.ai/46CPd58

Laura Weidinger, Jonathan Uesato, Maribeth Rauh, Conor Griffin, Po-Sen Huang, John Mellor, Amelia Glaese, Myra Cheng, Borja Balle, Atoosa Kasirzadeh, Courtney Biles, Sasha Brown, Zac Kenton, Will Hawkins, Tom Stepleton, Abeba Birhane, Lisa Anne Hendricks, Laura Rimell, William Isaac, Julia Haas, Sean Legassick, Geoffrey Irving, and Iason Gabriel (2022) Taxonomy of Risks posed by Language Models. In 2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT '22). Association for Computing Machinery, New York, NY, USA, 214–229. https://doi.org/10.1145/3531146.3533088

White House (2023a) Ensuring Safe, Secure, and Trustworthy AI. White House, https://www.whitehouse.gov/wp-content/uploads/2023/07/Ensuring-Safe-Secure-and-Trustworthy-AI.pdf

White House (2023b) Red-Teaming Large Language Models to Identify Novel AI Risks. White House, https://www.whitehouse.gov/ostp/news-updates/2023/08/29/red-teaming-large-language-models-to-identify-novel-ai-risks/

White House (2023c) Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence. White House, https://www.whitehouse.gov/briefing-room/presidential-actions/2023/10/30/executive-order-on-the-safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence/

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, Todor Mihaylov, Myle Ott, Sam Shleifer, Kurt Shuster, Daniel Simig, Punit Singh Koura, Anjali Sridhar, Tianlu Wang, and Luke Zettlemoyer (2022) OPT: Open Pre-trained Transformer Language Models. *arXiv*, https://arxiv.org/abs/2205.01068

Andy Zou, Zifan Wang, J. Zico Kolter, Matt Fredrikson (2023a) Universal and Transferable Adversarial Attacks on Aligned Language Models. *arXiv*, https://arxiv.org/abs/2307.15043

Andy Zou, Zifan Wang, J. Zico Kolter, Matt Fredrikson (2023b) LLM Attacks. https://github.com/llm-attacks/llm-attacks

Remco Zwetsloot and Allan Dafoe (2019) Thinking About Risks From AI: Accidents, Misuse and Structure. *Lawfare,* https://www.lawfareblog.com/thinking-about-risks-ai-accidents-misuse-and-structure


CLTC
Center for Long-Term
Cybersecurity
UC Berkeley