GPAI Evaluations Standards Taskforce: Towards Effective AI Governance [*†‡]

Patricia Paskov, RAND
ppaskov@rand.org

Lukas Berglund, RAND
lberglund@rand.org

Everett Smith, RAND
everetts@rand.org[§]

Lisa Soder, Interface
lsoder@interface-eu.org

Abstract

General-purpose AI (GPAI) evaluations have been proposed as a promising way of identifying and mitigating systemic risks posed by AI development and deployment. While GPAI evaluations play an increasingly central role in institutional decision- and policy-making – including by way of the European Union (EU) AI Act's mandate to conduct evaluations on GPAI models presenting systemic risk – no standards exist to date to promote their quality or legitimacy. To strengthen GPAI evaluations in the EU, which currently constitutes the first and only jurisdiction that mandates GPAI evaluations, we outline four desiderata for GPAI evaluations: internal validity, external validity, reproducibility, and portability. To uphold these desiderata in a dynamic environment of continuously evolving risks, we propose a dedicated EU GPAI Evaluation Standards Taskforce, to be housed within the bodies established by the EU AI Act. We outline the responsibilities of the Taskforce, specify the GPAI provider commitments that would facilitate Taskforce success, discuss the potential impact of the Taskforce on global AI governance, and address potential sources of failure that policymakers should heed.

1 Introduction

General-purpose AI (GPAI) evaluations have emerged as a valuable tool for measuring and addressing systemic risks posed by AI development and deployment. While GPAI evaluations play an increasingly central role in institutional decision- and policy-making – including by way of the European Union (EU) AI Act's mandate to conduct evaluations on GPAI models presenting systemic risk [22]– no standards exist to date to promote their quality or legitimacy. This paper

[*] The authors thank Anthony Barrett, Ben Bucknall, Herbie Bradley, Simeon Campos, Stephen Casper, Connor Dunlop, Jimmy Farrell, Arthur Goemans, Koen Holtman, David Manheim, Lorenzo Pacchiardi, Gabriel Sander, and Gaurav Sett for feedback and comments; and the Technology and Security Policy (TASP) leadership team, especially Michael Aird, Jeff Alstott, Chris Byrd, Casey Dugan, Ella Guest, and Brodi Kotila for their support.

[†] Funding for this work was provided by gifts from RAND supporters.

[‡] RAND working papers are intended to share early insights and solicit informal peer review. This working paper has been approved for circulation by RAND Global and Emerging Risks but has not yet completed formal peer review or been professionally edited or proofread. This working paper can be quoted and cited without permission of the author, provided the source is clearly referred to as a working paper. This working paper does not necessarily reflect the opinions of RAND's research clients and sponsors.

[§] Patricia Paskov, Lukas Berglund, and Everett Smith are Technology and Security Policy (TASP) fellows at RAND; for more information on the fellowship program, visit www.rand.org/tasp-fellows.

Socially Responsible Language Modelling Research (SoLaR) Workshop at NeurIPS 2024.

puts forth desiderata for GPAI evaluations and proposes a dedicated EU GPAI Evaluation Standards Taskforce to promote these desiderata adaptably in a dynamic environment of continuously evolving risks. This paper focuses specifically on an EU Taskforce, given the EU is, at the time of writing, the only jurisdiction with the power to mandate and regulate GPAI evaluation practices. A Taskforce in the EU, therefore, warrants timely consideration and enables direct policy influence. Actions in the EU now could plausibly pave the pathway for coordinated international standards in the future.

This paper is relevant to parties involved in or affected by decision-making processes surrounding GPAI evaluations, particularly in the EU. Primary audiences include the EU AI Office and the Codes of Practice working group chairs and participants. The paper's implications, however, reach beyond the EU, and towards international standards more broadly. Additional audiences may therefore include AI Safety Institutes (AISIs), civil society groups focused on AI governance, GPAI evaluation organisations, GPAI providers, independent researchers, and international standards organisations.

The paper proceeds as follows. Section 2 outlines the role of GPAI evaluations in risk assessment, risk mitigation, and AI governance, globally and in the EU. Section 3 calls for GPAI evaluations that follow four desiderata: internal validity, external validity, reproducibility, and portability. Section 4 highlights that the mere existence of standards to uphold these desiderata is in itself insufficient and potentially even harmful; instead, standards may best exist within an adaptive framework that can evolve with changes in technology, environment, and risk. Section 5 proposes the establishment of an EU GPAI Evaluations Standards Taskforce ("the Taskforce") to fulfill the needs highlighted in Section 3 and Section 4; and discusses the potential global impact of GPAI evaluations standard-setting in the EU. Section 6 concludes.

2 GPAI Risks and Evaluations

The development and deployment of general purpose AI (GPAI) models[5] present significant opportunity and considerable risks. Experts [32] and international leaders [1] alike have warned of systemic risks of AI systems [22, Recital 110], including:

- the misuse [6] of GPAI models, leading to increased scale and severity of cyber attacks, disinformation, and the development and use of chemical, biological, radiological, and nuclear (CBRN) weapons [10];
- the malfunctioning or misalignment [10] of GPAI models, resulting in loss of control, model autonomy [53], and deception; and
- broader societal risks related to democratic processes; public and economic security; the dissemination of illegal, false, or discriminatory content; environmental well-being, and inequality [79].

As such, experts have called for the global prioritisation of AI risk mitigation. As the scale and prevalence of AI systems multiply, this task becomes ever more critical.

GPAI evaluations play a key role in assessing and mitigating systemic risks [65]. Current GPAI evaluations [75], or the empirical assessment of the "components, capabilities, behaviour, and impact of an AI system," [77] include approaches like benchmarking [7], red-teaming [25, 55], and human uplift studies [49]. The results of such evaluations act as a proxy for GPAI model *capability* (how a GPAI model *could* behave in real-world deployment) and *alignment* (how GPAI model *would* behave in real-world deployment) [65].

Accordingly, GPAI evaluations are key in emerging AI governance processes. Evaluations form the foundation of major GPAI providers' scaling policies and safety frameworks[6], including those

[5] Synonymously referred to as foundation models. For the purposes of this paper, we use the term "GPAI models," consistent with the vocabulary of the EU AI Act.

[6] Anthropic's Responsible Scaling Policy, for example, outlines three AI Safety Levels (ASL), whereby model performance on evaluations determines the ASL of a model and the subsequent measures and safeguards to be implemented. DeepMind's Frontier Safety Framework establishes varying Critical Capability Levels, determined by model performance on model evaluations, at which a range of mitigations shall be triggered. OpenAI's Preparedness Framework defines model safety scores, as indicated by performance on evaluations, that correspond with key deployment and security decisions.

of Anthropic [3], DeepMind [18], and OpenAI [53]. Moreover, in Spring 2024, sixteen industry signatories, including GPAI providers like Amazon, Google, and Meta, committed to "consider[ing] results from internal and external evaluations…set[ting] out thresholds at which severe risks..would be intolerable" and "assess[ing] whether these thresholds have been breached," [23] the latter process in which GPAI evaluations would likely be essential.

From the government side, UK AI Safety Institute (AISI) identifies as one of its three core functions "develop[ing] and conduct[ing] evaluations on advanced AI systems," [34] and the US AISI aims to "champion the development of empirically grounded tests, benchmarks, and evaluations of AI models, systems, and agents" [35].

Uniquely, the European Union's (EU) AI Act mandates that providers must conduct evaluations on GPAI models presenting systemic risk [22, Article 55], making the EU, at the time of writing, the only jurisdiction with the power to mandate and regulate GPAI evaluation practices. While the AI Office has not yet determined the procedures for GPAI model evaluations [20], evaluation results could plausibly trigger regulatory actions, including, by the powers granted to the AI Office, requests for providers to implement safety measures and mitigations [22, Article 93.1a-b], submit additional information [22, Article 91], remove a model from the market ([22, Article 93.1c]; or, in extreme cases, the imposition of fines [22, Article 101]. As GPAI evaluations hold greater weight in decision- and policy-making – both within and across institutions – establishing and updating standards becomes increasingly important. This task is especially timely in the EU, the world's first-mover in government-mandated GPAI evaluations.

3 Desiderata for Governance-Enhancing GPAI Evaluations

Given their nascent and rapidly evolving state, GPAI evaluations are not currently beholden to standards or best practices [34, 4, 62]. To bolster AI governance and mitigate extreme risk within the EU and beyond, standards for GPAI evaluations may seek to uphold a core set of desiderata. This section proposes four such desiderata: internal validity, external validity, reproducibility, and portability. The proposed desiderata draw from evaluations literature and evaluation standards of other sectors[7], with a consideration of the unique characteristics of the GPAI evaluation ecosystem in the EU and beyond. The proposed desiderata, with closed-source GPAI models in mind, inherently foster transparency and can be even more readily achieved with open-source models. These desiderata should not be considered as exhaustive nor definitive, but rather as a starting point for ongoing discussion and refinement. This section defines these desiderata, outlines their policy relevance, surveys the status quo, and provides example standards that might uphold them.

3.1 Internal Validity

Internal validity refers to the extent to which the observed evaluation results represent the truth in the evaluation setting and are not due to methodological shortcomings [40, 43, 54]. Internal validity is crucial for providing consistent and unbiased insights into the probability and severity of threats presented by GPAI models.

In their current state, evaluations often lack internal validity due to confounding variables, measurement error, and a lack of robust statistical testing. Benchmark question phrasing [39, 47] and evaluation structure [63, 73], for example, have been shown to lead to substantially different results, suggesting that such results may be more reflective of methodological artifacts than of the true capabilities of GPAI models. Even holding the same phrasing and structure constant, results of a given evaluation may vary: Lukosiute [41], for example, shows that GPQA with chain-of-thought (CoT) prompting yields noisy results across ten runs. These findings underscore the need for well-considered sample sizes and well-powered statistical testing: in the absence of statistical power, meaningful capabilities may go undetected and, conversely, apparently significant capabilities may be exaggerated [14].

[7] For example, the European Union's Registration, Evaluation, Authorisation and Restriction of Chemicals (REACH) outlines regulations for chemical evaluations; and the International Standards Organisation (ISO) and International Electrotechnical Commission (IEC) specify voluntary evaluations standards for medical devices (ISO 13485), cyber and IT products (ISO/IEC 15408), construction materials (ISO 9001), and environmental management systems (ISO 14000).

Standards and practices that may promote internal validity are outlined in Appendix A.

3.2 External Validity

External validity refers to the extent to which evaluation results can be used as a proxy for model behavior in contexts outside of the evaluation environment. Within external validity lies construct validity, or "how well designed…the experimental setting is in relation to the research claim" [57]. External validity is key in bolstering confidence in the use of GPAI evaluations to trigger policy or regulatory decisions, especially in the context of a multi-stakeholder ecosystem [76].

An extensive body of research documents the ways in which GPAI evaluation results fail to hold under a range of real-world conditions: methods like scaffolding, fine-tuning, and prompting (e.g. single-turn, multi-turn, chain-of-thought) can meaningfully alter GPAI evaluation results [39, 74, 12, 3], yet no protocols guide the use and documentation of these methods. Furthermore, GPAI evaluation results often fail to serve as strong proxies for real-world risks. For example, while Meta evaluated Llama 3's safety by testing its propensity to comply with requests to output harmful content [44], its evaluations did not account for the removal of safeguards, which can be cheaply bypassed through fine-tuning [38]. As such, Meta's evaluation result cannot act as a strong proxy for the real-world risks posed by Llama 3.

Standards and practices that may promote external validity are outlined in Appendix A.

3.3 Reproducibility

Reproducibility refers to the ability to obtain consistent evaluation results using the same input data, computational methods, code, and evaluation conditions [51]. Reproducibility is key in bolstering confidence in the use of GPAI evaluations to trigger policy or regulatory decisions, especially in the context of a multi-stakeholder ecosystem [75]. Little attention has been given to date to the reproducibility of GPAI evaluations, though a need exists [26]. While reproducibility may not be applicable to all types of evaluations, its use should be strongly considered and its absence should be carefully justified.

Standards and practices that may promote reproducibility are outlined in Appendix A.

3.4 Portability

Portability refers to the ability of a range of stakeholders to consistently and seamlessly implement and assess GPAI evaluations across distinct institutions and hardware environments through, for example, accessible evaluation software. Portability facilitates both collaboration and external scrutiny, the latter of which can avail more reliable information by verifying GPAI provider claims and revealing new insights [2, 42]. Portability may furthermore enhance resource efficiency for governments and third parties, foster knowledge spillovers among stakeholders, and reduce pressure for industry concentration from regulation costs [67].

The EU AI Act [22]) and the Frontier AI Safety Commitments [23] highlight the advent of multi-stakeholder evaluation and, by extension, the need for portability. The EU AI Act outlines at least three distinct parties that may run GPAI evaluations: GPAI providers [22, Article 55.1a], the EU AI Office [22, Article 92.1], and qualified independent experts on behalf of the AI Office [22, Article 92.2]. Separately, signatories of the Frontier AI Safety Commitments commit to "consider[ing] results from internal and external evaluations as appropriate, such as by independent third-party evaluators, their home government, and other bodies their governments deem appropriate" [23]. Few common frameworks to date facilitate portability of evaluations. METR's Task Standard [45] standardizes task-based evaluations to facilitate efficiency, METR's Vivaria [46] provides a platform for running evaluations and conducting elicitation research, and the UK AI Safety Institute's Inspect [70] allows users to assess and score GPAI model capabilities in a range of areas, While these platforms offer a starting point for portability, substantial scope remains to promote more comprehensive and well-defined standards across evaluation types and contexts [61].

Standards and practices that may promote portability are outlined in Appendix A.

4 The Need for Adaptive Standards

Establishing standards for GPAI evaluations can promote desiderata like internal validity, external validity, reproducibility, and portability [77]. However, if overly rigid and static, standards may threaten the innovation and progress of GPAI evaluations. Given that GPAI evaluations are a nascent and rapidly evolving field in which best practices are subject to rapid technological and environmental advancements, the existence of standards in itself is insufficient and potentially even harmful if not paired with adaptivity [29].

Adaptivity refers to the ability of evaluation standards to anticipate and adjust nimbly to changes in technology, environment, and risk. GPAI evaluations are perpetually subject to change as landscapes shift, human interaction with models evolve, models gain new capabilities with potentially abrupt performance jumps, and new harm vectors emerge [24]. Furthermore, evaluations like benchmarks may become less reliable over time due to increasing likelihood of leakage into the training dataset [77]. Hence, in order to most effectively contribute to risk mitigation, GPAI evaluations methodology and measurement must continuously evolve.

5 GPAI Evaluations Standard Taskforce

To address the need for adaptive standards, we propose a GPAI Evaluations Standards Taskforce ("the Taskforce"), to be housed within institutions established by the EU AI Act. The Taskforce is a vetted body of independent researchers to maintain and update GPAI evaluation standards. Given the potential regulatory leverage of GPAI evaluations in the EU and the evolving nature of GPAI evaluations, a dedicated Taskforce can promote evaluation standards that foster desiderata like internal validity, external validity, reproducibility, and portability in an evolving manner in the EU and potentially beyond. This proposal follows a broader call for Audit Standards Boards [42] and proactive risk management in dynamic fields [60].

5.1 Institutional Setting

The provisions of the EU AI Act provide the scope for the Taskforce to exist [77]. Within the provisions of the EU AI Act, the Taskforce could sit within two potential institutional settings: The Scientific Panel of Experts (IS1) or The Advisory Forum (IS2). Appendix B provides details on these settings. Precedents for such a Taskforce within the EU exist and include that established by Commitment 27 of the 2022 Strengthened Code of Practice on Disinformation.[19]

5.2 Structure and Duties

The Taskforce could be composed of technical experts from academia, civil society, third-party model audit providers, regulators, and experts from government serving in a personal capacity, with Taskforce members furthermore fulfilling the criteria outlined in [22, Article 68.2], if falling within IS1, or [22, Article 67.2], if falling within IS2. The latter additionally allows for the involvement of GPAI industry experts. The Taskforce could facilitate regular pathways for input by a range of non-Taskforce members, including GPAI industry experts.[8]

Key Taskforce duties, drawing from the framework of the EU AI Act and building towards the desiderata highlighted in Section 3 are outlined in Appendix C.

5.3 GPAI Provider Commitments

The success of the Taskforce hinges upon the direct exchange with GPAI provider experts to support access and provide relevant information [77, 2]. The Frontier AI Safety Commitments[23] lay the groundwork for this collaboration and the EU AI Act Code of Practice[20], to be finalized by May 2025, could formally codify GPAI provider commitments to the Taskforce. To concretely support the development and ongoing success of the Taskforce, GPAI providers can commit to working with the relevant organizations (e.g. the European Commission, the EU AI Office, Civil Society, Data

[8] If falling within IS1, by way of [22, Article 68.2], Taskforce members must be independent from industry. Involving industry insights via non-Taskforce input is valuable.

Protection Authorities) to establish and sustain the Taskforce. In particular, this requires commitments to the provision of documentation; and of model and data access. Provider commitments are proposed in Appendix D.

5.4 Potential for Global Impact

The EU could plausibly pave the pathway for coordinated international standards of GPAI evaluations via a *de jure* Brussels effect (the adoption of EU standards by foreign governments) and a *de facto* Brussels effect (the unilateral regulation of GPAI provider practices) [66, 9]. Appendix E outlines these channels for global impact.

5.5 Potential Causes of Failure and Additional Approaches

While in the best-case scenario, the Taskforce could improve AI risk assessment and mitigation both in the EU and beyond, it could also face challenges. Appendix F outlines these challenges and offers related recommendations.

6 Conclusion

While GPAI evaluations have been proposed as a central tool in assessing and mitigating systemic risks posed by GPAI development and deployment, no established standards exist to date to promote their quality and legitimacy. This paper proposes an EU GPAI Evaluations Standards Taskforce to develop and adapt standards for desiderata of GPAI evaluations, including internal validity, external validity, reproducibility, and portability. Amidst the rapidly evolving state of GPAI evaluations and constant shifts in environments, technology, and risks, a Taskforce could promote relevant and governance-enhancing standards for effective risk assessment and mitigation in the EU and beyond. The success of the Taskforce and its responsibilities, as outlined in Appendix C, rely on collaboration with GPAI providers, as outlined in Appendix D. The impact of the Taskforce, for better or for worse, should not be underestimated: by way of *de facto* and *de jure* Brussels effects, the Taskforce could influence GPAI evaluations standards and GPAI safety at an international scale. As such, any design and management of the Taskforce should heed the potential pitfalls and recommendations highlighted in Appendix F.

This proposal for an EU GPAI Evaluation Standards Taskforce rests on four key assumptions: 1) well-formulated GPAI evaluations standards can support effective risk assessment, risk mitigation, and AI governance, 2) developing standards within institutions and environments capable of mandating and implementing standards increases the legitimacy and impact of those standards, 3) when establishing standards, the interactions fostered by a dedicated, multi-stakeholder Taskforce can ultimately increase the quality of standards, relative to many distinct, uncoordinated bi-lateral and multilateral interactions, and 4) the benefits to AI governance and public safety of establishing a GPAI Evaluations Standards Taskforce outweigh the costs.

This paper draws primarily on literature, current AI policy, and the authors' expertise. It faces the limitations endemic to the nascent field of GPAI evaluations and AI policy more broadly: sparse evidence, rapid advancements, and, accordingly, a lack of conclusive research from which to draw. Future research may employ qualitative methods like interviews to draw insights and recommendations from the multi-stakeholder GPAI evaluations community; as well as quantitative methods like surveys, impact evaluations, and forecasting to more rigorously estimate the impacts of an EU GPAI Evaluations Standards Taskforce on the state of risk assessment, risk mitigation, and AI governance in the EU and beyond.

Impact Statement

While GPAI evaluations are crucial for assessing and mitigating systemic risks, no standards currently exist to ensure their quality and legitimacy. This paper proposes an EU GPAI Evaluations Standards Taskforce to develop and adapt standards for robust, reproducible, and interoperable evaluations. The Taskforce, to be housed within institutions established by the EU AI Act, could enhance AI governance by providing adaptive frameworks for evaluations, on which policy-making processes increasingly rely. However, the success of the Taskforce may be stifled by the rapid pace of AI progress, potential bureaucratic obstacles, and the potential inadvertent stifling of safety-related innovation due to rigid standards. Despite these hurdles, establishing such a Taskforce could contribute to the responsible development and deployment of GPAI systems, both within and beyond the EU, and ultimately facilitate more effective governance in this critical domain.

Appendix A: Standards and Practices Promoting Desiderata

Internal validity

Standards and practices that may promote internal validity include:

- Specifications for hypothesis testing, sample sizes [40], random seeds, statistical significance, and statistical power [8];
- Specification and disclosure of environmental parameters used in evaluations [40];
- Controlling for confounding variables and mitigating spurious correlation (i.e. by randomising question order, ensuring balanced treatment and control groups in experimental settings, etc.); and
- Evidence of the absence of train-test contamination [37] and assurance that testing methodologies and test data are not used for GPAI system development.

External validity

Standards and practices that may promote results external validity include:

- The clear definition of risks to be measured;[9]
- Input of domain expertise in the design of evaluations, ensuring that the evaluation is a well-considered proxy for the broader real-world task or risk [40, 68];
- The use of strong elicitation in GPAI evaluations, including fine-tuning, scaffolding, white-box adversarial attacks [15], and prompt engineering; and
- Documentation of factors informative of the evaluation's external validity across contexts [27], including:
  - The environmental conditions in which the evaluation is run and the ways in which it converges with and diverges from the real-world context [40]; and
  - In the case of red-teaming and adversarial testing, details on the sourcing and vetting of domain experts to partake in the evaluation and the level and duration of access to GPAI models provided to these experts for the purpose of the evaluation.

Reproducibility

Standards and practices that may promote reproducibility include:

- Definition of environments and evaluations for which reproducibility can enhance risk assessment, risk mitigation, and AI governance;
- For the above-indicated environments and evaluations:
  - secure release of evaluation data, including input data, random seed, and output data for steps that are nondeterministic and cannot be reproduced [51]
  - secure release of evaluation code [28]; and

[9] See, for example, https://airisk.mit.edu/.

  - documentation of evaluation methodology, evaluation environment, computation al environment (i.e. hardware architecture, operating systemic, and library dependencies), elicitation methods, statistical testing, and analysis [52, 51].

Portability

Standards and practices that may promote portability include:

- The development and use of software that:
  - makes evaluations virtually "plug and play," reducing the engineering effort needed to implement another actor's evaluations;
  - is model- and agent-architecture-agnostic;[10]
  - facilitates swapping out and combining particular post-training enhancements between different evaluators; and
  - to address privacy constraints, where relevant, facilitates evaluation implementation in a privacy-preserving manner (e.g. without leaking additional information about a model or an evaluation to the other party) [5].

Appendix B: Institutional Setting

Institutional settings for the Taskforce include The Scientific Panel of Independent Experts (IS1) and The Advisory Forum (IS2):

IS1: The Scientific Panel of Independent Experts

The Scientific Panel of Independent Experts ("the Panel"), established by [22, Article 68 of the EU AI Act], is tasked with a range of responsibilities, including the "development of tools and methodologies for evaluating capabilities of general-purpose AI models and systems." Appendix B outlines the responsibilities of the Panel, notes Taskforce-relevant tasks, and highlights Panel duties that would be symbiotic with the proposed Taskforce work.

IS2: The Advisory Forum

The Advisory Forum ("the Forum"), established by [22, Article 67 of the EU AI Act], is tasked with providing technical expertise and advising the European Artificial Intelligence Board and the Commission. The Forum "may establish standing or temporary sub-groups as appropriate for the purpose of examining specific questions related to the objectives of this Regulation." By mandate, the Forum includes permanent advisory members from the European Committee for Standardization (CEN), which holds an technical cooperation agreement with the International Standards Organization (ISO) via the Vienna Agreement. This framework paves the path for potential dialogue between GPAI evaluations standards at the Taskforce-, EU-, and international-level.

Relative to alternative placements[11], IS1 and IS2 offer conditions particularly conducive to developing and implementing GPAI evaluation standards, including timely access to existing institutional infrastructure; the ability to convene expert pools that represent diverse, multi-stakeholder perspectives; regulatory ties to GPAI provider and models; the legal basis for implementation of GPAI evaluations and standards; and strong institutional ties to international-standard-setting organisations.

[10] e.g. evaluations that do not depend on a particular model or agent architecture, and can therefore be re-used with different models and agent architectures.

[11] Alternatives include, for example, the establishment of GPAI evaluations standards by third-party evaluators directly, GPAI providers directly, CEN/CENELEC, and the International Standards Organization (ISO).

Appendix C: Taskforce Integration in the Panel

Article 68.3[22], below, outlines the tasks of the Scientific Panel of Independent Experts. **T** demarcates tasks that could be partially or fully delegated to the Taskforce, given the tasks' direct relevance to GPAI evaluations standards. **TP** demarcates tasks that hold complementarities with GPAI evaluations but are not exhaustively fulfilled by the duties outlined in Section 5. Thus, the Taskforce could maintain a close dialogue with the Panel on these tasks, particularly on Task a.i., given its role in robustness and construct validity in GPAI evaluations. **P** demarcates tasks outside of the Taskforce's mandate.

> (a) supporting the implementation and enforcement of this Regulation as regards general-purpose AI models and systems, in particular by:
>
> (i) alerting the AI Office of possible systemic risks at Union level of GPAI models, in accordance with [22, Article 90] (**TP**);
>
> (ii) contributing to the development of tools and methodologies for evaluating capabilities of GPAI models and systems, including through benchmarks (**T**);
>
> (iii) providing advice on the classification of GPAI models with systemic risk (**P**);
>
> (iv) providing advice on the classification of various GPAI models and systems (**P**);
>
> (v) contributing to the development of tools and templates (**TP**);
>
> (b) supporting the work of market surveillance authorities, at their request (**P**);
>
> (c) supporting cross-border market surveillance activities as referred to in [22, Article 74.11], without prejudice to the powers of market surveillance authorities (**P**);
>
> (d) supporting the AI Office in carrying out its duties in the context of the Union safeguard procedure pursuant to [22, Article 81] (**P**).

Appendix C: Taskforce Duties

Harmonisation of Risk Taxonomies and GPAI Model Evaluation Methodologies

- Closely collaborating with the Scientific Panel of Independent Experts to promote a regularly updated comprehensive taxonomy and enumeration of GPAI systemic risks ("the Risk Taxonomy"), to be initially established by the Codes of Practice [22, Recital 116];
- Outlining and updating concrete examples of and standards for qualitative and quantitative experiments that would demonstrate the presence of the risk;
- Ensuring the continuous updating of and harmonisation between the Risk Taxonomy and GPAI evaluations standards; and
- Ensuring that updates to the taxonomy outline precise descriptions of risks, taking into consideration advancements in technology and the evolution of external environments, and drawing from resources including but not limited to provider safety cases reports, scaling risk management policies, capability forecast reports, and incident reporting documentation.

Standard-setting for GPAI Evaluations

- Regularly[12] publishing updated standards for GPAI model evaluations to guide evaluations implemented by providers [22, Article 55.1.a] and by the AI Office and third-party evaluators [22, Article 92] towards internal validity, external validity, reproducibility, and portability; and
- Defining environments and evaluations for which standards are not safety-enhancing.

Quality-control of Evaluations

- Setting standards for the third-party evaluator conduct, following the proposal in [58] for an audit oversight board modelled after the Public Company Accounting Oversight Board (PCAOB);

[12] Such as every twelve months or by qualified alert.

- Vetting, approving, and periodically[13] reviewing the fitness of third-party evaluators; and
- Auditing the methodologies and results of GPAI evaluations implemented by providers, the AI Office, and third-party evaluators both pre- and post-deployment.

Finally, while not a primary duty, the Taskforce should facilitate the harmonisation of GPAI model evaluation standards across GPAI providers and international bodies, including but not limited to standards organisations and AI Safety Institutes, where possible, provided such efforts do not compromise the independence of the Taskforce or the quality of its standards.

Appendix D: Provider Commitments

Documentation

A comprehensive understanding of evaluations results relies on intimate knowledge of the evaluations process [26][14]. Access to detailed documentation is crucial to account for nuances and potential pitfalls, interpret evaluations, and develop standards accordingly.

GPAI providers commit to draw-up and keep up-to-date sufficient documentation on their evaluation results and any additional information needed to rigorously audit these results. Sufficient documentation is understood to include, upon the determination of the AI office and proportional to the risk being assessed, the required methods and the security precautions taken by the Taskforce:

- Description of models and methods as outlined in [22, Annex XI.1], including but not limited to training data, training compute used, and number of perameters;[15]
- Documentation outlined in Annex XI.2.2.1 [22], including specifically:
  - Scores on individual benchmarks and other developers' evaluations;
  - Description of evaluations and benchmarks;
  - Model responses on benchmarks; and
  - Descriptions and examples of the data included in the benchmarks.
- Documentation outlined in Annex XI.2.2 [22], including specifically:
  - Description of capability elicitation procedures used in the benchmarks;
  - Calendar-time and employee-hours spent conducting the evaluation; and
  - The procedure used to fine-tune the model.
- Documentation outlined in Annex XI.2.3 [22], including specifically:
  - The scaffolding and tools available to the model during inference; and
  - Details about the model's system architecture.

GPAI providers additionally commit to providing updated documentation to the Taskforce, including but not limited to safety cases, scaling risk management policies, capability forecast reports, and incident reporting documentation, as outlined in Article 55.1.c and Recital 115. GPAI providers additionally commit to providing updated documentation to the Taskforce, including but not limited to safety cases, scaling risk management policies, capability forecast reports, and incident reporting documentation, as outlined in Article 55.1.c and Recital 115 [22].

[13] Such as every twelve months or by qualified alert

[14] Consider, for example, specialised evaluations like Bias Benchmark for QA (BBQ), which measure social biases and require nuanced interpretation. Human evaluations, while valuable for assessing real-world performance, are subject to variability based on evaluator characteristics and the inherent subjectivity of human judgment. Model-generated evaluations, while efficient, may inherit biases or inaccuracies from the models that created them.

[15] This information often includes valuable trade secrets, and providing it to outsiders risks leaking it to outsiders. On the other hand, knowing these details can be crucial for verifying safety cases and evaluation results. The risk associated with sharing this information should be weighed against the improved oversight that it enables.

Model and data access

Comprehensively implementing, drawing insights from, and setting standards for GPAI evaluations require sufficient model access. Sufficient access is understood to include, upon the determination of the Taskforce and proportional to the risk being assessed, the required methods and the security precautions taken by the Taskforce:

- The ability to use the GPAI system in the way that it will be made available to customers (e.g. prompting it for text or image responses);
- White box model access [15];
- Access to versions of the GPAI system that lack technical safety mitigations, as these can prevent evaluators from exploring the full range of an AI system's capabilities [71]
- Through an application programmer interface (API), the ability to fine-tune a GPAI system [69, 64, 13];
- Access to other models in the model family [13];
- Access to individual components of the GPAI system, including the core model (via API) and other software components, such as moderation filters, system prompts, and available plug-ins allowing additional capabilities like web browsing and code execution [13]; and
- The data used to train the GPAI model, and/or meta-data, descriptions, and examples of the training data.

Some forms of enhanced access may increase the risk that GPAI providers' intellectual property and trade secrets are leaked. To fulfil Article 78 of the EU AI Act [22], providers commit to working with the EU AI Office to provide sufficient access while minimising information security concerns. Techniques for achieving this objective may include:

- Structured API access, providing "de facto white-box" capabilities to third parties without giving direct access to model parameters [64];
- Physical solutions involving secure research environments on a GPAI provider's campus, allowing unrestricted white-box access while minimising risks of leaks;
- Legal mechanisms that draw on practices from other industries with audits can be used to enforce confidentiality and hold third parties accountable.

Appendix E: Brussels Effects

***De jure*, GPAI evaluations standards in the EU could inform norms and standards internationally by way of collaboration with AI Safety Institutes and/or via the CEN-ISO pipeline, as discussed in IS2 (Section 5)**. Evaluations may play a key policy role beyond the EU, including in the UK and US [34, 23]. As the potential first-mover in GPAI evaluations standards – and the only current jurisdiction able to mandate and regulate GPAI evaluation practices – the EU could plausibly pave the pathway for coordinated international standards. As AI Safety Institutes expand and synergise, the alignment of evaluation standards and the exchange of results could facilitate efficiency, returns to scale, and specialization [72].

***De facto*, GPAI evaluations standards could shape GPAI provider practices internationally.** A priori, anticipation and awareness of standards for GPAI evaluations mandated by the EU could preemptively shift GPAI provider behavior. Post-hoc, standards could impact evaluation results and trigger regulatory decisions. Both cases may, in turn, prompt changes in GPAI model development and deployment in the EU and potentially beyond. Previous estimates projected the EU's AI market share to be between 15% [33] - 22% [21]; depending on regulatory frameworks, this market share could provide strong incentives for GPAI providers to develop and deploy GPAI models that are provably safe, as measured by GPAI evaluations and accompanying standards.

Appendix F: Potential Causes of Failure and Additional Approaches

Beyond standard issues of funding, regulatory capture [17, 59, 78], and misaligned incentives [30], potential causes of failure for the Taskforce include:

- **Stifled acquisition of talent:** the Taskforce may face challenges in recruiting sufficiently talented experts. This could be exacerbated by an inability to match industry salaries or quality of life (even with adequate funding), restrictions on hiring industry-affiliated researchers, or other hiring process restrictions resulting in long hiring timelines or an inability to flexibly shape positions. Ensuring sufficient salary, nimble workstyle, and hiring flexibility could address these challenges [62].
- **Heavy bureaucratic processes:** while we emphasize adaptivity as a priority, in practice the Taskforce could struggle to keep pace with rapidly advancing AI capabilities. Bureaucratic processes, transparency and conflict of interest requirements and the need for consensus could slow down the general pace of work, restrict the ability of the Taskforce to gather information from experts and industry [36]. Granting the Taskforce the authority and political capital to establish its own processes from first principles instead of inherit processes from established industry players or government could address this.

In addition to the Taskforce merely failing to achieve its mission, two particularly prominent ways in which its efforts could backfire include:

- **Increased friction for evaluators:** standards will almost certainly create some extra processes for evaluators within GPAI providers as well as third parties. If this friction is too great, it could disincentivize new evaluators from entering the field at a time when talent is a critical bottleneck, and it could reduce the ability of evaluators to experiment with new methodologies and advance this critical risk management science[11].
- **False sense of security:** the existence of standards could create a false sense that AI risks are being adequately managed, potentially reducing vigilance or investment in other areas of risk management that might be more promising[50][56]. This is especially true insofar as it will be difficult to assess the effectiveness of these standards.

In light of the potential failures or downsides of the Taskforce, it is prudent to also consider other approaches to AI risk management, both as complements to and substitutes for the Taskforce, including but not limited to:

- **Outcome-based regulation:** focus on regulating the outcomes and impacts of AI systems rather than specific evaluation methodologies[16].
- **Mandatory information sharing:** require AI companies to share more detailed information about their systems and internal evaluation processes, fostering transparency, independent of specific standards[26].
- **Focus on infrastructure:** invest in shared testing infrastructure that can be used by multiple stakeholders to evaluate AI systems.
- **Bottom-up standards development:** encourage the development of standards through a distributed, community-driven process[48] involving a wide range of stakeholders, such as regularly posing open challenges or competitions to identify novel risks or evaluation methods, supplementing the work of a permanent Taskforce.
- **AI bounty programs:** establish programs that reward individuals or organisations for identifying potential risks or vulnerabilities in AI systems[31].

References

[1] AI Safety Summit. The Bletchley Declaration by Countries Attending the AI Safety Summit, 1-2 November 2023. 2023.

[2] Markus Anderljung, Everett Thornton Smith, Joe O'Brien, Lisa Soder, Benjamin Bucknall, Emma Bluemke, Jonas Schuett, Robert Trager, Lacey Strahm, and Rumman Chowdhury. Towards publicly accountable frontier LLMs: Building an external scrutiny ecosystem under the ASPIRE framework. *arXiv preprint arXiv:2311.14711*, 2023.

[3] Anthropic. Anthropic's responsible scaling policy. `https://www-cdn.anthropic.com/1adf000c8f675958c2ee23805d91aaade1cd4613/responsible-scaling-policy.pdf`, 2023. Accessed: September 6, 2024.

[4] Apollo Research. We need a science of evals. www.apolloresearch.ai/blog/we-need-a-science-of-evals, 2024. Accessed: September 6, 2024.

[5] Emma Bluemke, Tantum Collins, Ben Garfinkel, and Andrew Trask. Exploring the relevance of data privacy-enhancing technologies for ai governance use cases, 2023. URL https://arxiv.org/abs/2303.08956.

[6] Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021.

[7] Rishi Bommasani, Percy Liang, and Tony Lee. Holistic evaluation of language models. *Annals of the New York Academy of Sciences*, 1525(1):140–146, 2023.

[8] Samuel R Bowman and George E Dahl. What will it take to fix benchmarking in natural language understanding? *arXiv preprint arXiv:2104.02145*, 2021.

[9] Anu Bradford. *The Brussels effect: How the European Union rules the world*. Oxford University Press, USA, 2020.

[10] Miles Brundage, Shahar Avin, Jack Clark, Helen Toner, Peter Eckersley, Ben Garfinkel, Allan Dafoe, Paul Scharre, Thomas Zeitzoff, Bobby Filar, et al. The malicious use of artificial intelligence: Forecasting, prevention, and mitigation. *arXiv preprint arXiv:1802.07228*, 2018.

[11] Nils Brunsson, Andreas Rasche, and David Seidl. The dynamics of standardization: Three perspectives on standards in organization studies. *Organization Studies*, 33(5-6):613–632, 2012. doi: 10.1177/0170840612450120. URL https://doi.org/10.1177/0170840612450120.

[12] Sondos Mahmoud Bsharat, Aidar Myrzakhan, and Zhiqiang Shen. Principled instructions are all you need for questioning LLaMa-1/2, gpt-3.5/4. *arXiv preprint arXiv:2312.16171*, 2023.

[13] Benjamin S Bucknall and Robert F Trager. Structured Access for Third-Party Research on Frontier AI Models: Investigating Researchers' Model Access Requirements, 2023.

[14] Dallas Card, Peter Henderson, Urvashi Khandelwal, Robin Jia, Kyle Mahowald, and Dan Jurafsky. With little power comes great responsibility. *arXiv preprint arXiv:2010.06595*, 2020.

[15] Stephen Casper, Carson Ezell, Charlotte Siegmann, Noam Kolt, Taylor Lynn Curtis, Benjamin Bucknall, Andreas Haupt, Kevin Wei, Jérémy Scheurer, Marius Hobbhahn, et al. Black-box access is insufficient for rigorous AI audits. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*, pages 2254–2272, 2024.

[16] Cary Coglianese and Jennifer Nash. Performance-based regulation: Prospects and limitations in health, safety and environmental protection. *All Faculty Scholarship*, 2815, 2003.

[17] Ernesto Dal Bó. Regulatory capture: A review. *Oxford Review of Economic Policy*, 22(2): 203–225, 2006.

[18] Anca Dragan, Helen King, and Allan Dafoe. Frontier Safety Framework. https://storage.googleapis.com/deepmind-media/DeepMind.com/Blog/introducing-the-frontier-safety-framework/fsf-technical-report.pdf, 2024. Accessed: September 6, 2024.

[19] European Commission. 2022 Strengthened Code of Practice on Disinformation, 2022. URL https://digital-strategy.ec.europa.eu/en/library/2022-strengthened-code-practice-disinformation.

[20] European Commission. AI Act: Participate in the drawing-up of the first General-Purpose AI Code of Practice, 2024. URL https://digital-strategy.ec.europa.eu/en/news/ai-act-participate-drawing-first-general-purpose-ai-code-practice.

[21] European Commission, Directorate-General for Communications Networks, Content and Technology. Commission Staff Working Document Impact Assessment Accompanying the 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 SWD/2021/84 Final. Technical report, EUR-Lex, 2021. Accessed: September 6, 2024.

[22] European Parliament and Council of the European Union. Regulation of the European Parliament and of the Council laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act), May 2024. URL `https://data.consilium.europa.eu/doc/document/PE-24-2024-INIT/en/pdf`.

[23] Innovation Department for Science and Technology. Frontier AI Safety Commitments, AI Seoul Summit 2024. 2024. Accessed: September 6, 2024.

[24] Deep Ganguli, Danny Hernandez, Liane Lovitt, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Nova Dassarma, Dawn Drain, Nelson Elhage, et al. Predictability and surprise in large generative models. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, 2022.

[25] Deep Ganguli, Liane Lovitt, Jackson Kernion, Amanda Askell, Yuntao Bai, Saurav Kadavath, Ben Mann, Ethan Perez, Nicholas Schiefer, Kamal Ndousse, et al. Red teaming language models to reduce harms: Methods, scaling behaviors, and lessons learned. *arXiv preprint arXiv:2209.07858*, 2022.

[26] Deep Ganguli, Nicholas Schiefer, Marina Favaro, and Jack Clark. Challenges in Evaluating AI Systems. `https://www.anthropic.com/news/evaluating-ai-systems`, 2023. Accessed: September 6, 2024.

[27] Russell E Glasgow, Lawrence W Green, Lisa M Klesges, David B Abrams, Edwin B Fisher, Michael G Goldstein, Laura L Hayman, Judith K Ockene, and C Tracy Orleans. External validity: we need to do more. *Annals of Behavioral Medicine*, 31(2), 2006.

[28] Steven N Goodman, Daniele Fanelli, and John P A Ioannidis. What does research reproducibility mean? *Science Translational Medicine*, 8(341):341ps312–341ps312, 2016. doi: 10.1126/scitranslmed.aaf5027.

[29] Gudela Grote. Promoting safety by increasing uncertainty – implications for risk management. *Safety Science*, 2015. URL `https://doi.org/10.1016/j.ssci.2014.02.010`.

[30] Neel Guha, Christie Lawrence, Lindsey A Gailmard, Kit Rodolfa, Faiz Surani, Rishi Bommasani, Inioluwa Raji, Mariano-Florentino Cuéllar, Colleen Honigsberg, Percy Liang, et al. AI regulation has its own alignment problem: The technical and institutional feasibility of disclosure, registration, licensing, and auditing. *George Washington Law Review, Forthcoming*, 2023.

[31] Melissa Heikkila. A bias bounty for AI will help to catch unfair algorithms faster, 2022. Accessed: September 6, 2024.

[32] Geoffrey Hinton. Mitigating the risk of extinction from AI should be a global priority alongside other societal-scale risks such as pandemics and nuclear war, 2023. Statement on AI risk.

[33] IMF. European union: Share in global gross domestic product based on purchasing-power-parity from 2017 to 2027. Statista, 2022. Accessed: September 6, 2024.

[34] Innovation Department for Science and Technology. AI Safety Institute approach to evaluations. `https://www.gov.uk/government/publications/ai-safety-institute-approach-to-evaluations/ai-safety-institute-approach-to-evaluations`, 2024. Accessed: September 5, 2024.

[35] U.S. Artificial Intelligence Safety Institute. The United States Artificial Intelligence Safety Institute: Vision, Mission, and Strategic Goals. 2024.

[36] Wesley Kauffman, Gabel Taggart, and Barry Bozeman. Administrative delay, red tape, and organizational performance. pages 529–553, 2019. URL `https://doi.org/10.1080/15309576.2018.1474770`.

[37] Shachar Kaufman, Saharon Rosset, Claudia Perlich, and Ori Stitelman. Leakage in data mining: Formulation, detection, and avoidance. *ACM Trans. Knowl. Discov. Data*, 6(4), dec 2012. ISSN 1556-4681. doi: 10.1145/2382577.2382579. URL `https://doi.org/10.1145/2382577.2382579`.

[38] Simon Lermen, Charlie Rogers-Smith, and Jeffrey Ladish. Lora fine-tuning efficiently undoes safety training in llama 2-chat 70b. *arXiv preprint arXiv:2310.20624*, 2023.

[39] Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, et al. Holistic evaluation of language models. *arXiv preprint arXiv:2211.09110*, 2022.

[40] Thomas I. Liao, Rohan Taori, Inioluwa Deborah Raji, and Ludwig Schmidt. Are we learning yet? a meta-review of evaluation failures across machine learning, 2021.

[41] Kamilė Lukošiūtė. You need to be spending more money on evals. `https://kamilelukosiute.com/llms/You+need+to+be+spending+more+money+on+evals`, undated. Accessed: September 6, 2024.

[42] David Manheim, Sammy Martin, Mark Bailey, Mikhail Samin, and Ross Greutzmacher. The Necessity of AI Audit Standards Boards. *arXiv preprint arXiv:2404.13060*, 2024.

[43] Sandra Mathison. *Encyclopedia of Evaluation*. SAGE Publications, 2004.

[44] Meta. Model details. URL `https://github.com/meta-llama/llama3/blob/main/MODEL_CARD.md`.

[45] METR. Portable evaluation tasks via the metr task standard. `https://metr.org/blog/2024-02-29-metr-task-standard/`, 2024. Accessed: September 12, 2024.

[46] METR. Vivaria. `https://vivaria.metr.org/`, 2024. Accessed: September 12, 2024.

[47] Moran Mizrahi, Guy Kaplan, Dan Malkin, Rotem Dror, Dafna Shahaf, and Gabriel Stanovsky. State of what art? A call for multi-prompt LLM evaluation. *Transactions of the Association for Computational Linguistics*, 12:933–949, 2024.

[48] John B. Morris, Jr. Injecting the public interest into internet standards. 2011. doi: https://doi.org/10.7551/mitpress/8066.001.0001.

[49] Christopher A Mouton, Caleb Lucas, and Ella Guest. The Operational Risks of AI in Large-Scale Biological Attacks: Results of a Red-Team Study. Technical Report RR-A2977-2, RAND Corporation, 2024. URL `https://www.rand.org/pubs/research_reports/RRA2977-2.html`.

[50] Gabriel Mukobi. Reasons to Doubt the Impact of AI Risk Evaluations. 2024. URL `https://arxiv.org/pdf/2408.02565`.

[51] National Academies of Sciences, Engineering, and Medicine. *Reproducibility and Replicability in Science*. The National Academies Press, Washington, DC, 2019. ISBN 978-0-309-48616-3. doi: 10.17226/25303. URL `https://nap.nationalacademies.org/catalog/25303/reproducibility-and-replicability-in-science`.

[52] Brian A Nosek, Charles R Ebersole, Alexander C DeHaven, and David T Mellor. The preregistration revolution. *Proceedings of the National Academy of Sciences*, 115(11):2600–2606, 2018.

[53] OpenAI. Preparedness framework (beta), 2023. URL `https://cdn.openai.com/openai-preparedness-framework-beta.pdf`.

[54] Cecilia Maria Patino and Juliana Carvalho Ferreira. Internal and external validity: can you apply research study results to your patients? *J Bras Pneumol*, 44(3), 2018.

[55] Ethan Perez, Saffron Huang, Francis Song, Trevor Cai, Roman Ring, John Aslanides, Amelia Glaese, Nat McAleese, and Geoffrey Irving. Red teaming language models with language models. *arXiv preprint arXiv:2202.03286*, 2022.

[56] Michael Power. *Organized Uncertainty: Designing a World of Risk Management*. 2017.

[57] Inioluwa Deborah Raji, Emily M. Bender, Amandalynne Paullada, Emily Denton, and Alex Hanna. AI and the Everything in the Whole Wide World Benchmark. *arXiv preprint arXiv:2111.15366*, 2021. URL `https://arxiv.org/pdf/2111.15366`.

[58] Inioluwa Deborah Raji, Peggy Xu, Colleen Honigsberg, and Daniel Ho. Outsider oversight: Designing a third party audit ecosystem for ai governance. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, pages 557–571, 2022.

[59] Karthik Ramanna. *Political standards: Corporate interest, ideology, and leadership in the shaping of accounting rules for the market economy*. University of Chicago Press, 2015.

[60] Jens Rasmussen and Inge Suedung. *Proactive risk management in a dynamic society*. Swedish Rescue Services Agency, 2000. ISBN 9172530847.

[61] Anka Reuel, Ben Bucknall, Stephen Casper, Tim Fist, Lisa Soder, Onni Aarne, Lewis Hammond, Lujain Ibrahim, Alan Chan, Peter Wills, et al. Open problems in technical ai governance. *arXiv preprint arXiv:2407.14981*, 2024.

[62] Anka Reuel, Lisa Soder, Benjamin Bucknall, and Trond Arne Undheim. Position: Technical research and talent is needed for effective AI governance. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp, editors, *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 42543–42557. PMLR, 21–27 Jul 2024. URL `https://proceedings.mlr.press/v235/reuel24a.html`.

[63] Jaromir Savelka, Arav Agarwal, Christopher Bogart, and Majd Sakr. Large Language Models (GPT) Struggle to Answer Multiple-Choice Questions about Code. `https://arxiv.org/abs/2303.08033`, 2023. Accessed: October 17, 2024.

[64] Toby Shevlane. Structured access: an emerging paradigm for safe AI deployment. *arXiv preprint arXiv:2201.05159*, 2022.

[65] Toby Shevlane, Sebastian Farquhar, Ben Garfinkel, Mary Phuong, Jess Whittlestone, Jade Leung, Daniel Kokotajlo, Nahema Marchal, Markus Anderljung, Noam Kolt, et al. Model evaluation for extreme risks. *arXiv preprint arXiv:2305.15324*, 2023.

[66] Charlotte Siegmann and Markus Anderljung. The Brussels effect and artificial intelligence: How EU regulation will impact the global AI market. *arXiv preprint arXiv:2208.12645*, 2022.

[67] George Slover. Interoperability is important for competition, consumers, and the economy, 2023. URL `https://cdt.org/insights/interoperability-is-important-for-competition-consumers-the-economy/`.

[68] Elham Tabassi. *Artificial Intelligence Risk Management Framework*. NIST, 2023.

[69] Andrew Trask, Emma Bluemke, Ben Garfinkel, Claudia Ghezzou Cuervas-Mons, and Allan Dafoe. Beyond privacy trade-offs with structured transparency. *arXiv preprint arXiv:2012.08347*, 2020.

[70] UK AI Safety Institute. Inspect. `https://inspect.ai-safety-institute.org.uk/`, 2024. Accessed: September 12, 2024.

[71] UK Department of Science, Innovation, and Technology. Emerging Processes for Frontier AI Safety.

[72] Alexandre Variengien and Charles Martinet. AI Safety Institutes: Can countries meet the challenge? jul 2024. [Online; accessed CURRENT-DATE].

[73] Yixu Wang, Yan Teng, Kexin Huang, Chengqi Lyu, Songyang Zhang, Wenwei Zhang, Xingjun Ma, Yu-Gang Jiang, Yu Qiao, and Yingchun Wang. Fake alignment: Are llms really aligned well? `https://arxiv.org/abs/2311.05915s`, 2024. Accessed: October 17, 2024.

[74] Lucas Weber, Elia Bruni, and Dieuwke Hupkes. The icl consistency test. *arXiv preprint arXiv:2312.04945*, 2023.

[75] Laura Weidinger, Maribeth Rauh, Nahema Marchal, Arianna Manzini, Lisa Anne Hendricks, Juan Mateos-Garcia, Stevie Bergman, Jackie Kay, Conor Griffin, Ben Bariach, et al. Sociotechnical safety evaluation of generative AI systems. *arXiv preprint arXiv:2310.11986*, 2023.

[76] Laura Weidinger, Joslyn Barnhart, Jenny Brennan, Christina Butterfield, Susie Young, Will Hawkins, Lisa Anne Hendricks, Ramona Comanescu, Oscar Chang, and Mikel Rodriguez. Holistic safety and responsibility evaluations of advanced AI models. *arXiv preprint arXiv:2404.14068*, 2024.

[77] Laura Weidinger, Joslyn Barnhart, Jenny Brennan, Christina Butterfield, Susie Young, Will Hawkins, Lisa Anne Hendricks, Ramona Comanescu, Oscar Chang, Mikel Rodriguez, et al. Holistic safety and responsibility evaluations of advanced AI models. *arXiv preprint arXiv:2404.14068*, 2024.

[78] Tim Wu. In Regulating A.I., We May Be Doing Too Much. And Too Little., 11 2023. URL `https://www.nytimes.com/2023/11/07/opinion/biden-ai-regulation.html`. Accessed: 2024-01-23.

[79] Bengio Yohsua, Privitera Daniel, Besiroglu Tamay, Bommasani Rishi, Casper Stephen, Choi Yejin, Goldfarb Danielle, Heidari Hoda, Khalatbari Leila, Longpre Shayne, et al. International Scientific Report on the Safety of Advanced AI, 2024.