

Responsible Reporting for Frontier AI Development

Noam Kolt*^{1†}, Markus Anderljung², Joslyn Barnhart³, Asher Brass⁴,
Kevin Esvelt⁵, Gillian K. Hadfield^{1,6}, Lennart Heim², Mikel Rodriguez³,
Jonas B. Sandbrink⁷, Thomas Woodside⁸

¹University of Toronto

²Centre for the Governance of AI

³Google DeepMind

⁴Institute for AI Policy and Strategy

⁵Massachusetts Institute of Technology

⁶Vector Institute for AI

⁷University of Oxford

⁸Center for Security and Emerging Technology

noam.kolt@mail.utoronto.ca

Abstract

Mitigating the risks from frontier AI systems requires up-to-date and reliable information about those systems. Organizations that develop and deploy frontier systems have significant access to such information. By reporting safety-critical information to actors in government, industry, and civil society, these organizations could improve visibility into new and emerging risks posed by frontier systems. Equipped with this information, developers could make better informed decisions on risk management, while policymakers could design more targeted and robust regulatory infrastructure. We outline the key features of responsible reporting and propose mechanisms for implementing them in practice.

1 Introduction

Information is the lifeblood of good governance (Stephenson 2011; Van Loo 2019). Effectively responding to the new and emerging risks presented by frontier AI systems (Anderljung et al. 2023a; Toner et al. 2023) requires up-to-date and reliable information about those systems and their impact on society (Shevlane et al. 2023; Weidinger et al. 2023, 2024; Birhane et al. 2024). There is growing consensus among experts in AI safety and governance that reporting safety information to trusted actors in government and industry is key to achieving this goal (Bommasani et al. 2023a; Schuett et al. 2023; Kapoor et al. 2024; Woodside 2024). This is particularly the case for frontier models, i.e., highly capable foundation models that could pose severe risks to public safety (Anwar et al. 2024; Phuong et al. 2024).

Early efforts to facilitate reporting safety information made important strides. The AI Incident Database established by the Partnership on AI contains more than 2,000 reports of AI harms (McGregor 2021; McGregor, Paeth, and Lam 2022). The database, however, is limited to tangible

harms caused by deployed AI systems, as is the case for related initiatives (Turri and Dzombak 2023). These databases do not track anticipated risks, vulnerabilities, or near-misses (Frase and Hoffmann 2023; Shrishak 2023), and dedicate comparatively little attention to larger-scale or catastrophic risks (Bengio et al. 2024; Hendrycks, Mazeika, and Woodside 2023; Department for Science, Innovation and Technology, Foreign, Commonwealth & Development Office, and Prime Minister’s Office, 10 Downing Street 2023).

But the tide is changing. Recognizing the growing need to share information about AI safety with government actors, several leading developers committed to the U.S. government to “reporting their AI systems’ capabilities, limitations, and areas of appropriate and inappropriate use” and undertook to engage in “third-party discovery and reporting of vulnerabilities in their AI systems” (The White House 2023d). Some developers made additional commitments to share information with companies and governments (Frontier Model Forum 2023), including to provide actors in the UK government with “early or priority access to models for research and safety purposes to help build better evaluations” (Department for Science, Innovation and Technology et al. 2023). The UK government has also requested access to, and published, details concerning the safety practices of several leading AI companies (Department for Science, Innovation and Technology and AI Safety Institute 2023).

National governments and international institutions are also taking concrete steps to implement AI safety reporting. The European Union’s AI Act imposes stringent reporting obligations on the providers of high-risk AI systems (European Parliament 2024). An executive order issued by President Biden requires that AI developers provide the U.S. federal government with information regarding “activities related to training, developing, or producing dual-use foundation models,” as well as information regarding “the ownership and possession of the model weights” and the results of “red-team testing” (The White House 2023b). These requirements would initially apply to any model trained us-

*These authors contributed equally.

†Work done at Google DeepMind.

Copyright © 2024, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

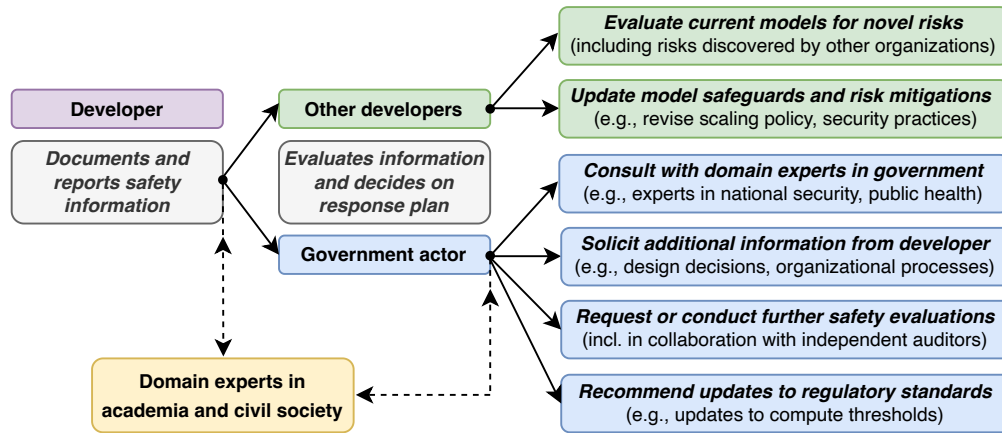


Figure 1: A framework for responsible reporting. Developers disclose safety-critical information to government actors and other developers, which decide on appropriate technical, organizational, and policy responses. Independent domain experts in academia and civil society receive key information and provide guidance to both developers and government actors.

ing more than 10^{26} operations, or any model using primarily genetic sequence data trained using 10^{23} operations. The OECD, meanwhile, has convened an expert group to develop an AI incident reporting framework (OECD.AI Network of Experts 2024).

Given this increasingly complex institutional context, distilling the key features of AI safety reporting is especially important. We aim to make headway on this challenge by clarifying the goals of reporting safety-related information (Section 2), describing the content of this information and to which actors it could be reported (Section 3), proposing institutional mechanisms to facilitate reporting (Section 4), and tackling potential hurdles to implementing these mechanisms in practice (Section 5). Taken together, these contributions complement and provide guidance for more concrete efforts to establish reporting frameworks, including multiple concurrent efforts being undertaken by actors in government, industry, and civil society.

2 Goals of Reporting

As in other industries with long-standing reporting practices, including healthcare, finance, and aviation (McGregor 2021; Raji et al. 2022; Turri and Dzombak 2023), information disclosures aim to achieve several goals. In the case of frontier AI systems, we focus on three main goals of reporting: (1) raising awareness among key stakeholders with regard to societal-scale impacts and risks from AI technologies; (2) incentivizing AI developers to adopt more robust risk management and safety practices; and (3) increasing regulatory visibility to enable policymakers to effectively respond to new risks, especially risks which government actors are best positioned to address.

2.1 Risk Awareness

In its simplest form, the case for reporting information about AI risks and vulnerabilities can be summarized as follows: "To make AI safer, we need to know when and how it fails"

(Arnold and Toner 2021). Access to such information is especially crucial in the case of frontier AI systems, whose risk profiles are continually changing due to their often unpredictable capabilities (Ganguli et al. 2022; Wei et al. 2022; Weidinger et al. 2023, 2024; Woodside 2024; Phuong et al. 2024).

Across both industry and government, "reporting builds a norm of admitting mistakes, noticing them, and sharing lessons learned" (Manheim 2023). Information about safety incidents and failures can offer valuable lessons on how and where risk mitigation measures fail (McGregor 2021). Alongside improving general awareness of these risks, both developers and government actors, supported by independent domain experts, need a detailed understanding of these risks and the methods for mitigating them. We focus on each in turn.

2.2 Industry Practices

Sharing information with other developers about risks from frontier AI systems and the measures for addressing them enables developers to design better risk management strategies and safety practices (Avin et al. 2021). For example, a developer may update its safety levels (Anthropic 2023),

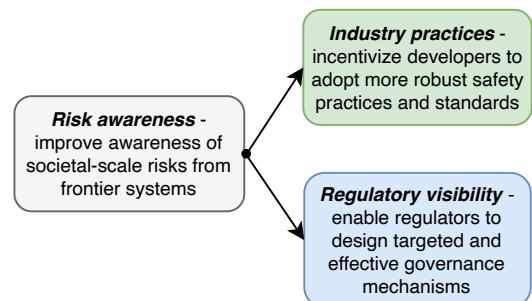


Figure 2: Goals of reporting safety information

preparedness framework (OpenAI 2023c), or capabilities scaling policy (Google DeepMind 2023) in response to information about new capabilities or risks (Model Evaluation and Threat Research 2023; Anderson-Samways et al. 2024). Information collected via reporting could also assist developers in establishing emergency response plans, internal audit procedures, and customer screening processes (Egan and Heim 2023; Schuett et al. 2023).

In addition to informing particular risk management practices, reporting could foster a stronger safety culture across the AI industry (Manheim 2023; Ojewale et al. 2024) and bring it closer in line with well-established reporting practices and norms in other industries, such as in the aviation industry (Federal Aviation Administration 2023b; Madsen, Dillon, and Tinsley 2015). For instance, regulation that mandates reporting the results of AI safety evaluations (The White House 2023b) could deter developers from accelerating development at the expense of safety (Askill, Brundage, and Hadfield 2019; Dafoe 2023). By enabling external actors and domain experts to verify the claims made by AI developers about the safety of their systems (Brundage et al. 2020; Avin et al. 2021; Sastry et al. 2024), reporting would subject frontier AI developers and their products to increased scrutiny. Consequently, less cautious actors would be incentivized to invest more in safety and adopt best practices employed by other organizations. As in relatively mature industries such as healthcare and finance, risk mitigation and safety could become an inherent and uncontroversial part of frontier AI development and deployment (Dobbe 2022; Hendrycks et al. 2022).

2.3 Regulatory Visibility

Reporting information about AI risks and potential mitigations is critical to informing the priorities and actions of policymakers. Without meaningful visibility into the technology’s design and use, policymakers cannot determine appropriate regulatory objectives, let alone build appropriate regulatory infrastructure (Anderl jung et al. 2023a; Hadfield, Cuéllar, and O’Reilly 2023). Accurate and up-to-date information about frontier AI systems and their impact is key to enabling policymakers to address the risks posed by these systems.

A reporting framework designed to furnish policymakers with safety-critical information will help address these concerns. Equipped with reliable and timely information about frontier AI systems, policymakers will be able to make better informed decisions about the goals and methods of regulation, and acquire the resources needed to take appropriate action (Whittlestone and Clark 2021; Clark 2023). For example, policymakers will be able to design or implement standards that are more responsive to trends in AI development and, ideally, preempt nascent and emerging risks (Kaminski 2023; Kolt 2024a; Arbel, Tokson, and Lin 2024). In particular, information collected from reporting will assist policymakers in tackling risks that government actors are best positioned to address, as observed by the UK government’s AI Safety Institute (Secretary of State for Science, Innovation and Technology 2023). For instance, upon receiving a report concerning national security risks (e.g., new

cyber capabilities or biological capabilities), national security experts in government could propose additional model evaluations or governance measures.

3 Decision-Relevant Information

The following section describes the categories of information that developers could report in the proposed framework, as well as the recipients of this information. The categories in Table 1 - *development and deployment, risks and harms, and mitigations* - are designed to provide government actors, developers, and independent domain experts with information that will assist in deciding on appropriate technical, organizational, and policy responses to novel AI capabilities and risks. In addition, the categories broadly align with recent regulatory regimes, including the disclosure requirements in the U.S. executive order (The White House 2023b), the EU AI Act (European Parliament 2024), and the UK proposal for AI risk reporting (Department for Science, Innovation and Technology 2023).

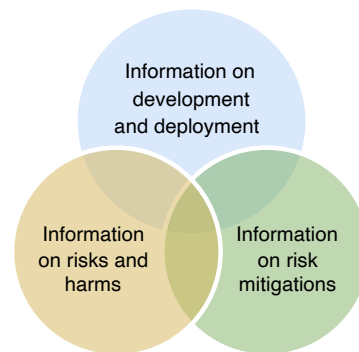


Figure 3: There is significant overlap between information pertaining to development and deployment and information pertaining to risks, harms, and mitigations.

Furnishing policymakers and domain experts with the above information is key to overcoming the inherent information deficit between industry and government (Karkkainen 2008). Despite repeated calls to provide governments with more comprehensive and consequential information relating to frontier AI technologies (Avin et al. 2021; Whittlestone and Clark 2021), regulators have often been caught off-guard, as exemplified by early drafts of the EU AI Act altogether failing to address foundation models.

As in other domains (Van Loo 2019), government actors need a deep understanding of the underlying technology, the resources required to build it, and the risks it may pose (Anderl jung et al. 2023a). For example, aviation regulators are authorized to conduct sweeping inspections of new aircraft technologies (e.g., (Federal Aviation Administration 2023a)), while financial regulators have privileged access to cutting-edge financial products and services in order to assess their anticipated impact on consumers and markets (e.g., (CFPB 2022)). Without comparable information on frontier AI systems (including information concerning development, risks, and mitigations), policymakers and do-




| Category | Content |
|---|---|
| Development and deployment  Recipients: Government actors | Details of state-of-the-art systems - copies of publicly available technical reports, including system and model cards, and additional information on training techniques, resources, and model capabilities. Information on current and upcoming training runs - description of architecture, compute, data collection, curation, filtering, and human feedback, training objectives (e.g., reward functions), and training techniques. Current and anticipated applications - description of the domains in which a model is currently deployed or anticipated to be deployed, the range of tasks they perform or are anticipated to perform, and usage trends and statistics. |
| Risks and harms  Recipients: Government actors, developers, and independent domain experts | Pre-deployment and post-deployment risk assessments - results of internal and external safety evaluations, including results of red-teaming and bounty programs. Concrete harms and safety incidents - description of incidents in which a system caused death or serious injury, damage to critical infrastructure, environmental harm, cybersecurity incidents, or other concrete harms, as well as harms that did not materialize (“near misses”). Dual-use and dangerous capabilities - evidence of a system exhibiting the ability to perform deception or manipulation, dual-use cyber capabilities or biological capabilities, weapons development, indications of the ability to engage in long-term planning, power-seeking, or other dangerous capabilities. |
| Mitigations  Recipients: Government actors, developers, and independent domain experts | Model alignment and safeguards - detailed explanation of alignment techniques, steps taken to prevent malicious use and other misuse (e.g., out-of-domain use), safety evaluations, and monitoring procedures. Organizational risk management - description of security standards, personnel and customer screening, auditing procedures, review processes, or other internal governance mechanisms, including circumstances in which such procedures were not effective or were not adopted. |

Table 1: Information categories, content, and recipients

main experts will be unable to assess for themselves the systems’ risk profiles or decide on the appropriate governance sites and mechanisms (Kolt 2024a,b).

Importantly, there is likely to be significant overlap between information pertaining to development and deployment (disclosed only to government actors) and information pertaining to risks and risk mitigations (disclosed to government actors, developers, and independent domain experts). Disentangling these two categories is not straightforward. For example, information regarding state-of-the art alignment techniques is both a model capability as well as a risk mitigation tool, making it unclear whether, or to what extent, such information should be disclosed to developers and independent domain experts, or only to government actors.

While we do not propose a precise definition distinguishing between the different categories, the appendices offer a concrete illustration of the kind of information that could fall into each category. Appendices A and B, which relate to cybersecurity and biosecurity, respectively, help shed light on which information would be disclosed to government actors only and which information would be disclosed to government actors, developers, and independent domain experts.

4 Institutional Framework

4.1 Contributors

Many different organizations are involved in developing and deploying frontier AI models (EPOCH 2024). These include organizations in industry, academia, and civil society, and across different geographies. While some organizations make their models accessible only via API (e.g., OpenAI, Anthropic), others publicly release the model weights subject to software licenses (e.g., Meta, EleutherAI).

Given that each of these organizations has expertise developing different models and deploying them in different contexts, each organization could offer valuable safety information. For example, an organization with extensive red-teaming experience could assist other organizations in designing protocols for external scrutiny of models (Anderljung et al. 2023b). Meanwhile, an organization that has developed methods to mitigate malicious use of its models could share those methods with other organizations.

Subject to the implementation challenges addressed below (Section 5), we suggest that all developers of frontier models could participate in the proposed framework and that the information they contribute could improve visibility into AI risks and mitigations.

4.2 Recipients

Government actors. As in disclosure regimes in other domains (McGregor 2021; Raji et al. 2022; Turri and Dzombak 2023), government actors are important recipients of the information provided under the proposed framework. Key characteristics for government actors include the following:

1. **Information security** - capacity to protect highly sensitive information and prevent its proliferation or misuse.
2. **Technical competence** - ability to understand, analyze, and draw conclusions from the information reported, including relevant domain expertise.
3. **Governance capacity** - organizational resources and legal authority to design and implement policy responses.
4. **Independence** - incentives and motivation to systematically and impartially execute policy responses.

In the United States, key actors include the U.S. Artificial Intelligence Safety Institute (U.S. Department of Commerce 2023, 2024), established through the National Institute of Standards and Technology (NIST), an agency that itself has significant in-house technical expertise and published an AI Risk Management Framework (National Institute of Standards and Technology 2021). Another key actor is the White House Office of Science and Technology Policy (OSTP), which released a Blueprint for an AI Bill of Rights (The White House 2022) and has been involved in facilitating model evaluations and securing voluntary commitments from leading AI developers (The White House 2023a,c,d). Other relevant actors include the National Security Council (NSC), Bureau of Industry and Security (BIS), Federal Trade Commission (FTC), and possibly new government bodies. In the United Kingdom, key actors include the UK government’s AI Safety Institute, which has indicated that it will work on conducting evaluations of advanced AI systems and facilitating information exchange (Department for Science, Innovation & Technology 2023).

Importantly, the combination of above characteristics is a new ‘muscle’ that governments will need to grow and flex. Effective reporting requires broad technical and sociotechnical capacity-building, which will require significant time and talent. In addition, it is worth noting that different government bodies might be better positioned to receive different types of information, instead of a single government body receiving all information disclosed under the framework. For example, dedicated cybersecurity agencies and biosecurity agencies might be the preferred recipients of information in their respective domains.

Developer reciprocity. As to which developers receive information under the reporting framework, we propose a principle of reciprocity according to which only developers that *contribute* information under the proposed framework will *receive* information under the framework. This principle both incentivizes developers to participate in the framework and prevents non-participating developers from free-riding. As illustrated in Figure 4 and Table 1, participating developers will only receive information relating to risks, harms, and mitigations, not commercially sensitive information re-

lating to model development and deployment - which will be disclosed to government actors only.

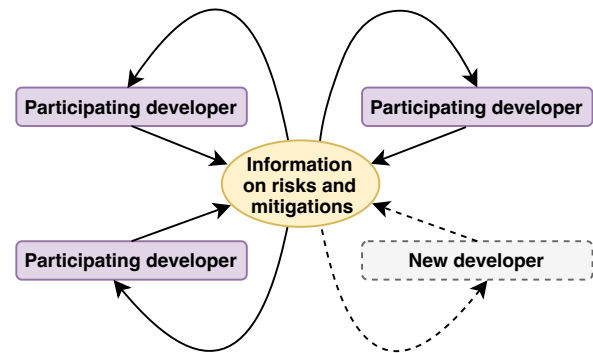


Figure 4: The principle of reciprocity incentivizes developers to join the responsible reporting framework by providing them with useful safety information.

Independent domain experts. As illustrated in Figure 1, independent domain experts in academia and civil society play two key roles in the proposed framework. First, domain experts will receive safety-critical information comparable to the information provided to participating developers. Second, domain experts will provide guidance to government actors with regard to the categories and content of information that developers report within the framework. In addition to the capacity to maintain highly effective information security, participating domain experts should have the following characteristics:

1. **Domain-specific expertise** - significant expertise in a risk-relevant field, especially a field or subfield in which government actors lack sufficient expertise.
2. **Security mindset** - ability and inclination to identify vulnerabilities, risks, and potential exploitations of AI systems and the environments in which they operate.

4.3 Documentation and Disclosure

Documentation. While there does not currently exist a dedicated government database for information relating to AI development and deployment (Clark 2023; Turri and Dzombak 2023), there are many mechanisms for documenting such information. These include data sheets (Gebu et al. 2021), model cards (Mitchell et al. 2019), reward reports (Gilbert et al. 2023), system cards (e.g., (OpenAI 2023a,b)), model reports (Gemini Team et al. 2023; Touvron et al. 2023; Reid et al. 2024), and ecosystem graphs (Bommasani et al. 2023b). Documentation may also need to be tailored to the particular model or system (Norval et al. 2022). For example, meaningful insight into models using reinforcement learning from human feedback (RLHF) can only be achieved via granular information on the relevant human feedback, reward model, and policy (Casper et al. 2023; Lambert, Gilbert, and Zick 2023).

Drawing on these and other documentation practices, we offer a preliminary picture of how developers could record

and communicate information under the proposed framework - as illustrated in Table 2. Notably, leading developers already report some of this information (including the developers of GPT-4, Claude 3, Llama 2, and Gemini).

| Information categories | Documentation |
|-----------------------------------|--|
| Development and deployment | Model cards for current models and models under development, including details of data collection, model training and finetuning, evaluation metrics, intended uses, ecosystem dependencies (e.g., compute sponsors, API access), and model maintenance plan. |
| Risks and harms | Results of pre-deployment and post-deployment safety evaluations (including underlying code and data), records of all safety incidents (including actors involved, causes of incidents, and responses to incident), and additional threat intelligence and impact assessments. |
| Mitigations | Description of technical safeguards for preventing misuse and accident risks (including alignment and monitoring procedures) and organizational measures (e.g., documents establishing internal safety and governance structures). |

Table 2: Developer documentation for each category of safety information

Although major jurisdictions increasingly mandate reporting, the precise scope and form of reporting remain open to interpretation. For example, the U.S. executive order requires that developers provide “information, reports, or records” regarding “training, developing, or producing dual-use foundation models” and “the ownership and possession of the model weights”, but does not specify precisely what information those reports must contain (The White House 2023b). Nor does the executive order prescribe the exact form in which information should be documented, leaving these fine-grained, yet important, tasks to future policy instruments.

In addition, given that developers’ safety evaluations and reporting practices are likely to evolve over time, policymakers will need to continually refine their governance responses to information received, which will require ongoing investment in both technical competence and institutional capacity (see 4.2).

Disclosure process. In addition to carefully documenting safety information, developers will need to securely communicate that information to the intended recipients. Doing so is critical to preventing the information from being intercepted or exploited by malicious actors, especially in the

case of information relating to dangerous dual-use capabilities. The high-level security practices that developers have been advised to adopt internally (e.g., (Schuett et al. 2023)) should apply to external information disclosures as well.

A further issue concerns the circumstances of disclosure, that is, the stages of design, development, and deployment in which developers report the above information. For example, should the circumstances of disclosure be determined by reference to a particular time (e.g. n days or weeks prior to and/or following deployment) or to a certain capabilities or risk threshold (which may be harder to define)? Should the timing and frequency of disclosure differ between reporting information to government actors compared with reporting information to other developers or domain experts? To operationalize responsible reporting, we need sufficiently flexible, yet clear, answers to these questions.

5 Implementation

5.1 Challenges

There are several challenges to successfully implementing the proposed framework for responsible reporting. We divide these challenges into two categories: (a) challenges facing developers that seek to participate in the reporting framework and (b) broader challenges concerning the overall effectiveness of the framework.

Challenges for developers. Developers seeking to engage in responsible reporting are likely to confront four main challenges.

1. **Intellectual property.** Commercially sensitive information (e.g., descriptions of new models or capabilities, logs of real-world incidents) could be inadvertently disclosed to, or exploited by, other developers and competitors.
2. **Reputational risk.** Reporting safety incidents and anticipated risks or vulnerabilities could damage a developer’s reputation and harm their business interests.
3. **Legal liability.** Disclosing certain safety information could potentially increase developers’ exposure to legal liability. Moreover, a legal obligation to disclose the results of safety tests may deter some developers from conducting rigorous safety tests in the first place.
4. **Coordination among developers.** Developers may be reluctant to participate in the framework and be exposed to business risks without assurances that their competitors will also participate and be similarly exposed to such risks.

Broader challenges. The potential obstacles to the framework achieving its goals (set out in Section 2) can be grouped into four broad challenges.

1. **Evaluation, documentation, and reporting resources.** Developers may lack the resources to effectively collect, document, and report the information required by the proposed framework.
2. **Misreporting.** Developers may inadvertently or deliberately report information that is either inaccurate or incomplete, undermining its reliability and usefulness.

3. **Information hazards.** Information reported under the framework could be intercepted by malicious actors and used for nefarious purposes, or be misused by its intended recipients.
4. **Institutional capacity.** Actors that receive information under the framework may lack the capacity to protect, analyze, or effectively respond to the information provided.

For further discussion of these and other challenges facing disclosure mechanisms, see (Guha et al. 2024), including concerns relating to firm-level and broader compliance costs, the impact of disclosure on design choices in AI development, and the potential for disclosures adversely impacting governance decisions.

5.2 Pathways Forward

Some of the above challenges could be addressed through targeted institutional mechanisms, some of which are already incorporated in the proposed framework. Other challenges require broader structural intervention. In this section, we assess how to address the most salient concerns along two different pathways.

Voluntary Implementation

The first pathway involves implementing responsible reporting as part of a voluntary governance regime (McGregor 2021; Marchant and Gutierrez 2022; Bommasani et al. 2023a). Developers voluntarily commit to partake in the reporting framework, whether in the absence of, or alongside, a broader regulatory regime. In this scenario, we propose the following institutional mechanisms:

1. **Differential disclosure.** Concerns regarding the protection of intellectual property and commercially sensitive information are largely addressed by features of the framework already discussed. Developers disclose information pertaining to model development and deployment to government actors only, not to competitors or other developers. For example, a safety incident report would describe the hazardous use observed, but would not disclose the relevant model’s architecture, training methods, compute, or data. In addition, developers could differentially disclose information *within* government. For example, developers might disclose information about dangerous dual-use capabilities to some (rather than all) participating government actors.
2. **Anonymized reporting.** To protect developers’ reputations, certain potentially damaging information disclosed under the framework could be de-identified, such that it could not be attributed to a particular developer and would not tarnish their reputation (Brundage et al. 2020; Turri and Dzombak 2023). Notably, effective anonymization may be difficult to achieve in some circumstances, such as where the developer identity can be inferred from the evaluations conducted. Anonymization is probably more appropriate for reporting information to participating developers and domain experts, not government actors. Government actors that demonstrate the characteristics set out above (4.2), including reliable information

security, should receive de-anonymized versions of the information disclosed under the framework.

3. **Organizational pre-commitments.** Developers could collectively commit in advance to participate in the reporting framework. Such commitments could be supported by a bond-like regime in which developers make upfront payments (prior to joining the framework) that are incrementally returned to developers contingent on their good faith participation in the framework.

Regulatory Implementation

The second pathway involves integrating responsible reporting into a broader purpose-built regulatory regime, such as the U.S. executive order (The White House 2023b) or EU AI Act (European Parliament 2024). In this scenario, we suggest the following institutional mechanisms will help facilitate more informative and actionable reporting:

1. **Liability safe harbors.** Developers’ reluctance to disclose information that may increase their legal exposure could be tackled by regulation that introduces safe harbor provisions that protect companies from legal liability arising from participation in the reporting framework (AI Policy and Governance Working Group 2023; Longpre et al. 2024). These could be modeled on existing safe harbors in environmental regulation and financial regulation.
2. **Government resourcing.** Under a purpose-built regulatory regime, government actors could be allocated resources to develop the technical and governance capacity to protect, analyze, and effectively respond to information disclosed under the framework (e.g., (Department for Science, Innovation & Technology 2023)). Equipped with these resources, government actors could also assist developers in making the required disclosures.
3. **Enforcement.** If regulations imposed legal sanctions in the event of negligent or deliberate misreporting, developers would be strongly incentivized to establish organizational processes for ensuring good faith and effective reporting. Independent auditors approved by regulators could also assist in detecting misreporting (Falco et al. 2021; Mökander et al. 2023; Casper et al. 2024).

6 Conclusion

Improvements in AI safety and governance hinge on the information available to key stakeholders. Building on existing efforts in government, industry, and civil society, responsible reporting aims to facilitate communicating and responding to safety-critical information in a dedicated secure institutional framework. While the implementation of responsible reporting faces several challenges, there are promising pathways forward. Frontier developers could begin by voluntarily reporting information about risks and mitigations that goes beyond current regulatory requirements. Policy-makers, meanwhile, could integrate features of responsible reporting into emerging governance regimes.

Appendices

A Cybersecurity




| Category | Key items |
|---|--|
| <p>Development and deployment</p> <p> Recipients: Government actors only</p> | <ol style="list-style-type: none"> 1. List of training datasets and descriptions of data sanitization or anonymization practices. 2. List of training-related hardware (including processing units, networking hardware, and peripheral equipment). 3. Software bill of material (SBOM) for training runs. 4. List of known vulnerabilities in software/hardware components, products, and libraries used in model development and/or deployment. 5. List of mitigations and/or justification for using products despite known vulnerabilities. 6. List of cloud providers, resources, and physical data center locations for training runs. 7. Is the model intended for use in offensive cyber operations and/or defensive cybersecurity activities? 8. Is the model intended for use in software, firmware, hardware, or cryptographic development? |
| <p>Risks and harms</p> <p> Recipients: Government actors, developers, and independent domain experts</p> | <ol style="list-style-type: none"> 1. Results of: (a) external and internal penetration tests, safety evaluations, and vulnerability scans; (b) bounty programs and disclosed vulnerabilities/exploits; (c) threat modeling assessments. 2. Description of incidents in which a system: (a) discovered novel vulnerabilities in software, firmware, hardware, or cryptographic products; (b) developed malware utilized in illicit activities; (c) was used to illegally access private data, gain access to an unauthorized network, or exfiltrate sensitive information from a device or network. 3. Complete incident reports for: (a) external breaches and unauthorized access of model weights or other training infrastructure; (b) insider or third party leaking of model weights or other training infrastructure. 4. Evidence of a system exhibiting the ability to perform: (a) vulnerability and exploit discovery (including static/dynamic code analysis, protocol reverse-analysis, and vulnerability to exploit conversion); (b) malware development and deployment; (c) social engineering attacks (e.g., phishing); (d) compromise of cryptographic systems or protocols; (e) model self-replication. 5. Report of cybersecurity-related capability evaluations. |
| <p>Mitigations</p> <p> Recipients: Government actors, developers, and independent domain experts</p> | <ol style="list-style-type: none"> 1. Description of security standards, personnel screening, and auditing procedures, including: (a) encryption standards for data at rest and in transit; (b) digital access controls (e.g., RBAC); (c) physical access controls; (d) patch management; (e) versioning controls; (f) backup integrity and verification. 2. Technical description of how users can interact with the model, including: (a) API specifications, security standards, and auditing practices; (b) credential provisioning and security practices. |

Table 3: Examples of cybersecurity information (including cyber capabilities and security vulnerabilities) that developers could report as part of the proposed framework

B Biosecurity

| Category | Key items |
|---|--|
| <p>Development and deployment</p> <p> Recipients: Government actors only</p> | <ol style="list-style-type: none"> 1. List of all biology-related data used in training, including papers, experimental protocols, and datasets relating to any life sciences field. 2. Description of methods to optimize for particular biological capabilities (e.g., host-pathogen interaction prediction, genetic sequence analysis or assembly, or structural outputs) or call specific biological tools. 3. Evaluation of biological science capabilities (such as conceptual ideation, experimental design, knowledge pooling and teaching, laboratory standard operating procedures and tacit knowledge, and sequence design capabilities). 4. Description of intended user base and deployment strategy (e.g., laypeople or a particular research or practitioner community, API design and deployment). |
| <p>Risks and harms</p> <p> Recipients: Government actors, developers, and independent domain experts</p> | <ol style="list-style-type: none"> 1. Description of biorisk capability evaluations and red-teaming, including: (a) evaluation methodologies and information hazard risk mitigations; (b) team size, types of participants (including independent domain experts), and skillsets (including security mindset); (c) details of scaffolding and finetuning methods (including specialized biology or chemistry tools). 2. Results of biorisk capability evaluations, especially evidence of dual-use capabilities and marginal improvements of AI models over existing (non-AI) methods, including the ability to: (a) provide dual-use biological information; (b) describe which biological agents or constructs are most hazardous and accessible; (c) instruct how to acquire or synthesize a controlled agent or a pandemic pathogen; (d) perform end-to-end synthesis of a controlled agent; (e) create a viable alternative structure for a controlled agent or enhancements mapping onto experiments of concern; (f) ideate novel biological tools by combining concepts or natural capabilities; (g) assist in the weaponization of biology; (h) exhibit evidence of security mindset with respect to biological vulnerabilities. 3. Records of the use of biorisk capabilities, including: (a) access to dual-use biorisk capabilities; (b) violations of model usage policies. |
| <p>Mitigations</p> <p> Recipients: Government actors, developers, and independent domain experts</p> | <ol style="list-style-type: none"> 1. Description of measures to reduce biorisk capabilities and harmful outputs, including: (a) decisions regarding the inclusion (or non-inclusion) in training data of papers, experimental protocols, and datasets relating to dual-use biology; (b) finetuning and other methods that can cause a model to refrain from performing certain biorisk-related tasks (e.g., accessing and delivering controlled agents and potential pandemic pathogens); (c) preventing jailbreaks and other adversarial attacks; (d) decisions regarding the ability (or inability) of models to call third-party biological tools that have not been evaluated for dangerous biological capabilities during assessment. 2. Monitoring and controlling model usage: (a) collecting know-your-customer (KYC) information on users who seek to access certain dual-use biorisk capabilities (e.g., developing novel pathogens); (b) restricting access to these capabilities, including plans to adhere to the principle of least privilege (PoLP) through user access controls. 3. Establishment of biosecurity incident, threat alert, and escalation processes, including: (a) investigation procedures in the event users violate usage policies; (b) incident reporting mechanisms for unanticipated post-deployment demonstration of dangerous biological capabilities; (c) response plan to deliberate malicious use of advanced biological capabilities. |

Table 4: Examples of biosecurity information that developers could report as part of the proposed framework

Acknowledgements

For helpful comments and suggestions, we thank Conor Griffin, Lewis Ho, Séb Krier, Lucy Lim, Aalok Mehta, Nikhil Mulani, Cassidy Nelson, Cullen O’Keefe, Sophie Rose, Yonadav Shavit, and Toby Shevlane.

References

- AI Policy and Governance Working Group. 2023. Comment of the AI Policy and Governance Working Group on the NTIA AI Accountability Policy Request for Comment Docket NTIA-230407-0093.
- Anderljung, M.; Barnhart, J.; Korinek, A.; Leung, J.; O’Keefe, C.; Whittlestone, J.; Avin, S.; Brundage, M.; Bullock, J.; Cass-Beggs, D.; Chang, B.; Collins, T.; Fist, T.; Hadfield, G.; Hayes, A.; Ho, L.; Hooker, S.; Horvitz, E.; Kolt, N.; Schuett, J.; Shavit, Y.; Siddarth, D.; Trager, R.; and Wolf, K. 2023a. Frontier AI Regulation: Managing Emerging Risks to Public Safety. ArXiv:2307.03718 [cs].
- Anderljung, M.; Smith, E. T.; O’Brien, J.; Soder, L.; Bucknall, B.; Bluemke, E.; Schuett, J.; Trager, R.; Strahm, L.; and Chowdhury, R. 2023b. Towards Publicly Accountable Frontier LLMs: Building an External Scrutiny Ecosystem under the ASPIRE Framework. ArXiv:2311.14711 [cs].
- Anderson-Samways, B.; Ee, S.; O’Brien, J.; Buhl, M.; and Williams, Z. 2024. Responsible Scaling: Comparing Government Guidance and Company Policy. Technical report, Institute for AI Policy and Strategy.
- Anthropic. 2023. Anthropic’s Responsible Scaling Policy.
- Anwar, U.; Saparov, A.; Rando, J.; Paleka, D.; Turpin, M.; Hase, P.; Lubana, E. S.; Jenner, E.; Casper, S.; Sourbut, O.; Edelman, B. L.; Zhang, Z.; Günther, M.; Korinek, A.; Hernandez-Orallo, J.; Hammond, L.; Bigelow, E.; Pan, A.; Langosco, L.; Korbak, T.; Zhang, H.; Zhong, R.; hÉigeartaigh, S. Ó.; Recchia, G.; Corsi, G.; Chan, A.; Anderljung, M.; Edwards, L.; Bengio, Y.; Chen, D.; Albanie, S.; Maharaj, T.; Foerster, J.; Tramer, F.; He, H.; Kasirzadeh, A.; Choi, Y.; and Krueger, D. 2024. Foundational Challenges in Assuring Alignment and Safety of Large Language Models. arXiv:2404.09932.
- Arbel, Y. A.; Tokson, M.; and Lin, A. 2024. Systemic Regulation of Artificial Intelligence. *Arizona State Law Journal*, forthcoming.
- Arnold, Z.; and Toner, H. 2021. AI Accidents: An Emerging Threat. Technical report, CSET.
- Askill, A.; Brundage, M.; and Hadfield, G. 2019. The Role of Cooperation in Responsible AI Development. ArXiv:1907.04534 [cs].
- Avin, S.; Belfield, H.; Brundage, M.; Krueger, G.; Wang, J.; Weller, A.; Anderljung, M.; Krawczuk, I.; Krueger, D.; Lebensold, J.; Maharaj, T.; and Zilberman, N. 2021. Filling gaps in trustworthy development of AI. *Science*, 374(6573): 1327–1329. Publisher: American Association for the Advancement of Science.
- Bengio, Y.; Hinton, G.; Yao, A.; Song, D.; Abbeel, P.; Darrell, T.; Harari, Y. N.; Zhang, Y.-Q.; Xue, L.; Shalev-Shwartz, S.; Hadfield, G.; Clune, J.; Maharaj, T.; Hutter, F.; Baydin, A. G.; McIlraith, S.; Gao, Q.; Acharya, A.; Krueger, D.; Dragan, A.; Torr, P.; Russell, S.; Kahneman, D.; Brauner, J.; and Mindermann, S. 2024. Managing Extreme AI Risks amid Rapid Progress. *Science*, 384(6698): 842–845.
- Birhane, A.; Steed, R.; Ojewale, V.; Vecchione, B.; and Raji, I. D. 2024. AI auditing: The Broken Bus on the Road to AI Accountability. ArXiv:2401.14462 [cs].
- Bommasani, R.; Klyman, K.; Longpre, S.; Kapoor, S.; Maslej, N.; Xiong, B.; Zhang, D.; and Liang, P. 2023a. The Foundation Model Transparency Index. ArXiv:2310.12941 [cs].
- Bommasani, R.; Soylu, D.; Liao, T. I.; Creel, K. A.; and Liang, P. 2023b. Ecosystem Graphs: The Social Footprint of Foundation Models. ArXiv:2303.15772 [cs].
- Brundage, M.; Avin, S.; Wang, J.; Belfield, H.; Krueger, G.; Hadfield, G.; Khlaaf, H.; Yang, J.; Toner, H.; Fong, R.; Maharaj, T.; Koh, P. W.; Hooker, S.; Leung, J.; Trask, A.; Bluemke, E.; Lebensold, J.; O’Keefe, C.; Koren, M.; Ryffel, T.; Rubinovitz, J. B.; Besiroglu, T.; Carugati, F.; Clark, J.; Eckersley, P.; de Haas, S.; Johnson, M.; Laurie, B.; Ingerman, A.; Krawczuk, I.; Askill, A.; Cammarota, R.; Lohn, A.; Krueger, D.; Stix, C.; Henderson, P.; Graham, L.; Prunkl, C.; Martin, B.; Seger, E.; Zilberman, N.; hÉigeartaigh, S. O.; Kroeger, F.; Sastry, G.; Kagan, R.; Weller, A.; Tse, B.; Barnes, E.; Dafoe, A.; Scharre, P.; Herbert-Voss, A.; Rasser, M.; Sodhani, S.; Flynn, C.; Gilbert, T. K.; Dyer, L.; Khan, S.; Bengio, Y.; and Anderljung, M. 2020. Toward Trustworthy AI Development: Mechanisms for Supporting Verifiable Claims. ArXiv:2004.07213 [cs].
- Casper, S.; Davies, X.; Shi, C.; Gilbert, T. K.; Scheurer, J.; Rando, J.; Freedman, R.; Korbak, T.; Lindner, D.; Freire, P.; Wang, T. T.; Marks, S.; Segerie, C.-R.; Carroll, M.; Peng, A.; Christoffersen, P.; Damani, M.; Slocum, S.; Anwar, U.; Siththaranjan, A.; Nadeau, M.; Michaud, E. J.; Pfau, J.; Krashennikov, D.; Chen, X.; Langosco, L.; Hase, P.; Biyik, E.; Dragan, A.; Krueger, D.; Sadigh, D.; and Hadfield-Menell, D. 2023. Open Problems and Fundamental Limitations of Reinforcement Learning from Human Feedback. *Transactions on Machine Learning Research*.
- Casper, S.; Ezell, C.; Siegmann, C.; Kolt, N.; Curtis, T. L.; Bucknall, B.; Haupt, A.; Wei, K.; Scheurer, J.; Hobbhahn, M.; Sharkey, L.; Krishna, S.; Von Hagen, M.; Alberti, S.; Chan, A.; Sun, Q.; Gerovitch, M.; Bau, D.; Tegmark, M.; Krueger, D.; and Hadfield-Menell, D. 2024. Black-Box Access Is Insufficient for Rigorous AI Audits. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, FAccT ’24, 2254–2272. New York, NY, USA: Association for Computing Machinery. ISBN 9798400704505.
- CFPB. 2022. CFPB Supervision and Examination Process.
- Clark, J. 2023. Information Markets and AI Development. In Bullock, J. B.; Chen, Y.-C.; Himmelreich, J.; Hudson, V. M.; Korinek, A.; Young, M. M.; and Zhang, B., eds., *The Oxford Handbook of AI Governance*. Oxford University Press, 1 edition. ISBN 978-0-19-757932-9.
- Dafoe, A. 2023. AI Governance: Overview and Theoretical Lenses. In Bullock, J. B.; Chen, Y.-C.; Himmelreich,

- J.; Hudson, V. M.; Korinek, A.; Young, M. M.; and Zhang, B., eds., *The Oxford Handbook of AI Governance*. Oxford University Press. ISBN 978-0-19-757932-9.
- Department for Science, Innovation & Technology. 2023. Introducing the AI Safety Institute.
- Department for Science, Innovation and Technology. 2023. Policy paper: Emerging processes for frontier AI safety. Technical report, GOV.UK.
- Department for Science, Innovation and Technology; and AI Safety Institute. 2023. Policy Updates.
- Department for Science, Innovation and Technology; AI Safety Institute; Chloe Smith MP; and The Rt Hon Rishi Sunak MP. 2023. Tech entrepreneur Ian Hogarth to lead UK's AI Foundation Model Taskforce.
- Department for Science, Innovation and Technology; Foreign, Commonwealth & Development Office; and Prime Minister's Office, 10 Downing Street. 2023. The Bletchley Declaration by Countries Attending the AI Safety Summit, 1-2 November 2023.
- Dobbe, R. I. J. 2022. System Safety and Artificial Intelligence. In Bullock, J. B.; Chen, Y.-C.; Himmelreich, J.; Hudson, V. M.; Korinek, A.; Young, M. M.; and Zhang, B., eds., *The Oxford Handbook of AI Governance*. Oxford University Press, 1 edition. ISBN 978-0-19-757932-9.
- Egan, J.; and Heim, L. 2023. Oversight for Frontier AI through a Know-Your-Customer Scheme for Compute Providers. ArXiv:2310.13625 [cs].
- EPOCH. 2024. Epoch Database.
- European Parliament. 2024. Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act) Text with EEA relevance.
- Falco, G.; Shneiderman, B.; Badger, J.; Carrier, R.; Dabhura, A.; Danks, D.; Eling, M.; Goodloe, A.; Gupta, J.; Hart, C.; Jirotko, M.; Johnson, H.; LaPointe, C.; Llorens, A. J.; Mackworth, A. K.; Maple, C.; Pálsson, S. E.; Pasquale, F.; Winfield, A.; and Yeong, Z. K. 2021. Governing AI safety through independent audits. *Nature Machine Intelligence*, 3(7): 566–571. Publisher: Nature Publishing Group.
- Federal Aviation Administration. 2023a. The Inspection Process.
- Federal Aviation Administration. 2023b. Mandatory and Voluntary Incident Reporting.
- Frase, H.; and Hoffmann, M. 2023. Adding Structure to AI Harm. Technical report, CSET.
- Frontier Model Forum. 2023. Frontier Model Forum: Advancing frontier AI safety.
- Ganguli, D.; Hernandez, D.; Lovitt, L.; Askell, A.; Bai, Y.; Chen, A.; Conerly, T.; Dassarma, N.; Drain, D.; Elhage, N.; El Showk, S.; Fort, S.; Hatfield-Dodds, Z.; Henighan, T.; Johnston, S.; Jones, A.; Joseph, N.; Kernian, J.; Kravec, S.; Mann, B.; Nanda, N.; Ndousse, K.; Olsson, C.; Amodei, D.; Brown, T.; Kaplan, J.; McCandlish, S.; Olah, C.; Amodei, D.; and Clark, J. 2022. Predictability and Surprise in Large Generative Models. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency, FAccT '22, 1747–1764*. New York, NY, USA: Association for Computing Machinery. ISBN 978-1-4503-9352-2.
- Gebbru, T.; Morgenstern, J.; Vecchione, B.; Vaughan, J. W.; Wallach, H.; Daumé III, H.; and Crawford, K. 2021. Datasheets for Datasets. ArXiv:1803.09010 [cs].
- Gemini Team; Anil, R.; Borgeaud, S.; Wu, Y.; Alayrac, J.-B.; Yu, J.; Soricut, R.; Schalkwyk, J.; Dai, A. M.; Hauth, A.; Millican, K.; Silver, D.; Petrov, S.; Johnson, M.; Antonoglou, I.; Schrittwieser, J.; Glaese, A.; Chen, J.; Pitler, E.; Lillicrap, T.; Lazaridou, A.; Firat, O.; Molloy, J.; Isard, M.; Barham, P. R.; Hennigan, T.; Lee, B.; Viola, F.; Reynolds, M.; Xu, Y.; Doherty, R.; Collins, E.; Meyer, C.; Rutherford, E.; Moreira, E.; Ayoub, K.; Goel, M.; Tucker, G.; Piqueras, E.; Krikun, M.; Barr, I.; Savinov, N.; Danihelka, I.; Roelofs, B.; White, A.; Andreassen, A.; von Glehn, T.; Yagati, L.; Kazemi, M.; Gonzalez, L.; Khalman, M.; Sygnowski, J.; Frechette, A.; Smith, C.; Culp, L.; Proleev, L.; Luan, Y.; Chen, X.; Lottes, J.; Schucher, N.; Lebron, F.; Rustemi, A.; Clay, N.; Crone, P.; Kocisky, T.; Zhao, J.; Perz, B.; Yu, D.; Howard, H.; Bloniarz, A.; Rae, J. W.; Lu, H.; Sifre, L.; Maggioni, M.; Alcober, F.; Garrette, D.; Barnes, M.; Thakoor, S.; Austin, J.; Barth-Maron, G.; Wong, W.; Joshi, R.; Chaabouni, R.; Fatiha, D.; Ahuja, A.; Liu, R.; Li, Y.; Cogan, S.; Chen, J.; Jia, C.; Gu, C.; Zhang, Q.; Grimsstad, J.; Hartman, A. J.; Chadwick, M.; Tomar, G. S.; Garcia, X.; Senter, E.; Taropa, E.; Pillai, T. S.; Devlin, J.; Laskin, M.; Casas, D. d. L.; Valter, D.; Tao, C.; Blanco, L.; Badia, A. P.; Reitter, D.; Chen, M.; Brennan, J.; Rivera, C.; Brin, S.; Iqbal, S.; Surita, G.; Labanowski, J.; Rao, A.; Winkler, S.; Parisotto, E.; Gu, Y.; Olszewska, K.; Zhang, Y.; Ad-danki, R.; Miech, A.; Louis, A.; Shafey, L. E.; Teplyashin, D.; Brown, G.; Catt, E.; Attaluri, N.; Balaguer, J.; Xiang, J.; Wang, P.; Ashwood, Z.; Briukhov, A.; Webson, A.; Ganapathy, S.; Sanghavi, S.; Kannan, A.; Chang, M.-W.; Stjerngren, A.; Djolonga, J.; Sun, Y.; Bapna, A.; Aitchison, M.; Pejman, P.; Michalewski, H.; Yu, T.; Wang, C.; Love, J.; Ahn, J.; Bloxwich, D.; Han, K.; Humphreys, P.; Sellam, T.; Bradbury, J.; Godbole, V.; Samangooei, S.; Damoc, B.; Kaskasoli, A.; Arnold, S. M. R.; Vasudevan, V.; Agrawal, S.; Riesa, J.; Lepikhin, D.; Tanburn, R.; Srinivasan, S.; Lim, H.; Hodkinson, S.; Shyam, P.; Ferret, J.; Hand, S.; Garg, A.; Paine, T. L.; Li, J.; Li, Y.; Giang, M.; Neitz, A.; Abbas, Z.; York, S.; Reid, M.; Cole, E.; Chowdhery, A.; Das, D.; Rogozińska, D.; Nikolaev, V.; Sprechmann, P.; Nado, Z.; Zilka, L.; Prost, F.; He, L.; Monteiro, M.; Mishra, G.; Welty, C.; Newlan, J.; Jia, D.; Allamanis, M.; Hu, C. H.; de Liedekerke, R.; Gilmer, J.; Saroufim, C.; Rijhwani, S.; Hou, S.; Shrivastava, D.; Baddepudi, A.; Goldin, A.; Oztirel, A.; Cassirer, A.; Xu, Y.; Sohn, D.; Sachan, D.; Amplayo, R. K.; Swanson, C.; Petrova, D.; Narayan, S.; Guez, A.; Brahma, S.; Landon, J.; Patel, M.; Zhao, R.; Villela, K.; Wang, L.; Jia, W.; Rahtz, M.; Giménez, M.; Yeung, L.; Lin, H.; Keeling, J.; Georgiev, P.; Mincu, D.; Wu, B.; Haykal, S.; Saputro, R.; Vodrahalli, K.; Qin, J.; Cankara, Z.; Sharma,

A.; Fernando, N.; Hawkins, W.; Neyshabur, B.; Kim, S.; Hutter, A.; Agrawal, P.; Castro-Ros, A.; Driessche, G. v. d.; Wang, T.; Yang, F.; Chang, S.-y.; Komarek, P.; McIlroy, R.; Lučić, M.; Zhang, G.; Farhan, W.; Sharman, M.; Natsev, P.; Michel, P.; Cheng, Y.; Bansal, Y.; Qiao, S.; Cao, K.; Shakeri, S.; Butterfield, C.; Chung, J.; Rubenstein, P. K.; Agrawal, S.; Mensch, A.; Soparkar, K.; Lenc, K.; Chung, T.; Pope, A.; Maggiore, L.; Kay, J.; Jhakra, P.; Wang, S.; Maynez, J.; Phuong, M.; Tobin, T.; Tacchetti, A.; Trebacz, M.; Robinson, K.; Katariya, Y.; Riedel, S.; Bailey, P.; Xiao, K.; Ghelani, N.; Aroyo, L.; Slone, A.; Houlsby, N.; Xiong, X.; Yang, Z.; Gribovskaya, E.; Adler, J.; Wirth, M.; Lee, L.; Li, M.; Kagohara, T.; Pavagadhi, J.; Bridgers, S.; Bortsova, A.; Ghemawat, S.; Ahmed, Z.; Liu, T.; Powell, R.; Bolina, V.; Iinuma, M.; Zablotskaia, P.; Besley, J.; Chung, D.-W.; Dozat, T.; Comanescu, R.; Si, X.; Greer, J.; Su, G.; Polacek, M.; Kaufman, R. L.; Tokumine, S.; Hu, H.; Buchatskaya, E.; Miao, Y.; Elhawaty, M.; Siddhant, A.; Tomasev, N.; Xing, J.; Greer, C.; Miller, H.; Ashraf, S.; Roy, A.; Zhang, Z.; Ma, A.; Filos, A.; Besta, M.; Blevins, R.; Klimenko, T.; Yeh, C.-K.; Changpinyo, S.; Mu, J.; Chang, O.; Pajarskas, M.; Muir, C.; Cohen, V.; Lan, C. L.; Haridasan, K.; Marathe, A.; Hansen, S.; Douglas, S.; Samuel, R.; Wang, M.; Austin, S.; Lan, C.; Jiang, J.; Chiu, J.; Lorenzo, J. A.; Sjöstrand, L. L.; Cevey, S.; Gleicher, Z.; Avrahami, T.; Boral, A.; Srinivasan, H.; Selo, V.; May, R.; Aisopos, K.; Hussienot, L.; Soares, L. B.; Baumli, K.; Chang, M. B.; Recasens, A.; Caine, B.; Pritzel, A.; Pavetic, F.; Pardo, F.; Gergely, A.; Frye, J.; Ramasesh, V.; Horgan, D.; Badola, K.; Kassner, N.; Roy, S.; Dyer, E.; Campos, V.; Tomala, A.; Tang, Y.; Badawy, D. E.; White, E.; Mustafa, B.; Lang, O.; Jindal, A.; Vikram, S.; Gong, Z.; Caelles, S.; Hemsley, R.; Thornton, G.; Feng, F.; Stokowiec, W.; Zheng, C.; Thacker, P.; Ünlü, Ç.; Zhang, Z.; Saleh, M.; Svensson, J.; Bileschi, M.; Patil, P.; Anand, A.; Ring, R.; Tsihlias, K.; Vezer, A.; Selvi, M.; Shevlane, T.; Rodriguez, M.; Kwiatkowski, T.; Daruki, S.; Rong, K.; Dafoe, A.; FitzGerald, N.; Gu-Lemberg, K.; Khan, M.; Hendricks, L. A.; Pellat, M.; Feinberg, V.; Cobon-Kerr, J.; Sainath, T.; Rauh, M.; Hashemi, S. H.; Ives, R.; Hasson, Y.; Li, Y.; Noland, E.; Cao, Y.; Byrd, N.; Hou, L.; Wang, Q.; Sottiaux, T.; Paganini, M.; Lespiau, J.-B.; Moufarek, A.; Hassan, S.; Shivakumar, K.; van Amersfoort, J.; Mandhane, A.; Joshi, P.; Goyal, A.; Tung, M.; Brock, A.; Sheahan, H.; Misra, V.; Li, C.; Rakićević, N.; Dehghani, M.; Liu, F.; Mittal, S.; Oh, J.; Noury, S.; Sezener, E.; Huot, F.; Lamm, M.; De Cao, N.; Chen, C.; Elsayed, G.; Chi, E.; Mahdieh, M.; Tenney, I.; Hua, N.; Petrychenko, I.; Kane, P.; Scandinaro, D.; Jain, R.; Uesato, J.; Datta, R.; Sadosky, A.; Bunyan, O.; Rabbiej, D.; Wu, S.; Zhang, J.; Vasudevan, G.; Leurent, E.; Al-nahlawi, M.; Georgescu, I.; Wei, N.; Zheng, I.; Chan, B.; Rabinovitch, P. G.; Stanczyk, P.; Zhang, Y.; Steiner, D.; Naskar, S.; Azzam, M.; Johnson, M.; Paszke, A.; Chiu, C.-C.; Elias, J. S.; Mohiuddin, A.; Muhammad, F.; Miao, J.; Lee, A.; Vieillard, N.; Potluri, S.; Park, J.; Davoodi, E.; Zhang, J.; Stanway, J.; Garmon, D.; Karmarkar, A.; Dong, Z.; Lee, J.; Kumar, A.; Zhou, L.; Evens, J.; Isaac, W.; Chen, Z.; Jia, J.; Levskaya, A.; Zhu, Z.; Gorgolewski, C.; Grabowski, P.; Mao, Y.; Magni, A.; Yao, K.; Snaider, J.; Casagrande, N.; Suganthan, P.; Palmer, E.; Irving, G.; Loper, E.; Faruqi, M.; Arkatkar, I.; Chen, N.; Shafran, I.; Fink, M.; Castaño, A.; Giannoumis, I.; Kim, W.; Rybiński, M.; Sreevatsa, A.; Prendki, J.; Soergel, D.; Goedeckemeyer, A.; Gierke, W.; Jafari, M.; Gaba, M.; Wiesner, J.; Wright, D. G.; Wei, Y.; Vashisht, H.; Kulizhskaya, Y.; Hoover, J.; Le, M.; Li, L.; Iwuanyanwu, C.; Liu, L.; Ramirez, K.; Khorlin, A.; Cui, A.; LIN, T.; Georgiev, M.; Wu, M.; Aguilar, R.; Pallo, K.; Chakladar, A.; Repina, A.; Wu, X.; van der Weide, T.; Ponnappalli, P.; Kaplan, C.; Simsa, J.; Li, S.; Dousse, O.; Yang, F.; Piper, J.; Ie, N.; Lui, M.; Pasumarthi, R.; Lintz, N.; Vijayakumar, A.; Thiet, L. N.; Andor, D.; Valenzuela, P.; Paduraru, C.; Peng, D.; Lee, K.; Zhang, S.; Greene, S.; Nguyen, D. D.; Kurylowicz, P.; Velury, S.; Krause, S.; Hardin, C.; Dixon, L.; Janzer, L.; Choo, K.; Feng, Z.; Zhang, B.; Singhal, A.; Latkar, T.; Zhang, M.; Le, Q.; Abellan, E. A.; Du, D.; McKinnon, D.; Antropova, N.; Bolukbasi, T.; Keller, O.; Reid, D.; Finchelstein, D.; Raad, M. A.; Crocker, R.; Hawkins, P.; Dadashi, R.; Gaffney, C.; Lall, S.; Franko, K.; Filonov, E.; Bulanova, A.; Leblond, R.; Yadav, V.; Chung, S.; Askham, H.; Cobo, L. C.; Xu, K.; Fischer, F.; Xu, J.; Sorokin, C.; Alberti, C.; Lin, C.-C.; Evans, C.; Zhou, H.; Dimitriev, A.; Forbes, H.; Banarse, D.; Tung, Z.; Liu, J.; Omernick, M.; Bishop, C.; Kumar, C.; Sterneck, R.; Foley, R.; Jain, R.; Mishra, S.; Xia, J.; Bos, T.; Cideron, G.; Amid, E.; Piccinno, F.; Wang, X.; Banzal, P.; Gurita, P.; Noga, H.; Shah, P.; Mankowitz, D. J.; Polozov, A.; Kushman, N.; Krakovna, V.; Brown, S.; Bateni, M.; Duan, D.; Firoiu, V.; Thotakuri, M.; Natan, T.; Mohananey, A.; Geist, M.; Mudgal, S.; Girgin, S.; Li, H.; Ye, J.; Roval, O.; Tojo, R.; Kwong, M.; Lee-Thorp, J.; Yew, C.; Yuan, Q.; Bagri, S.; Sinopalnikov, D.; Ramos, S.; Mellor, J.; Sharma, A.; Severyn, A.; Lai, J.; Wu, K.; Cheng, H.-T.; Miller, D.; Sonnerat, N.; Vnukov, D.; Greig, R.; Beattie, J.; Caveness, E.; Bai, L.; Eischenschlos, J.; Korchemniy, A.; Tsai, T.; Jasarevic, M.; Kong, W.; Dao, P.; Zheng, Z.; Liu, F.; Yang, F.; Zhu, R.; Geller, M.; Teh, T. H.; Sanmiya, J.; Gladchenko, E.; Trdin, N.; Sozanschi, A.; Toyama, D.; Rosen, E.; Tavakkol, S.; Xue, L.; Elkind, C.; Woodman, O.; Carpenter, J.; Papamakarios, G.; Kemp, R.; Kaffe, S.; Grunina, T.; Sinha, R.; Talbert, A.; Goyal, A.; Wu, D.; Owusu-Afriyie, D.; Du, C.; Thornton, C.; Pont-Tuset, J.; Narayana, P.; Li, J.; Fatehi, S.; Wieting, J.; Ajmeri, O.; Uria, B.; Zhu, T.; Ko, Y.; Knight, L.; Héliou, A.; Niu, N.; Gu, S.; Pang, C.; Tran, D.; Li, Y.; Levine, N.; Stolovich, A.; Kalb, N.; Santamaria-Fernandez, R.; Goenka, S.; Yustalim, W.; Strudel, R.; Elqursh, A.; Lakshminarayanan, B.; Deck, C.; Upadhyay, S.; Lee, H.; Dusenberry, M.; Li, Z.; Wang, X.; Levin, K.; Hoffmann, R.; Holtmann-Rice, D.; Bachem, O.; Yue, S.; Arora, S.; Malmi, E.; Mirylenka, D.; Tan, Q.; Koh, C.; Yeganeh, S. H.; Pöder, S.; Zheng, S.; Pongetti, F.; Tariq, M.; Sun, Y.; Ionita, L.; Seyedhosseini, M.; Tafti, P.; Kotikalapudi, R.; Liu, Z.; Gulati, A.; Liu, J.; Ye, X.; Chrzaszcz, B.; Wang, L.; Sethi, N.; Li, T.; Brown, B.; Singh, S.; Fan, W.; Parisi, A.; Stanton, J.; Kuang, C.; Koverkathu, V.; Choquette-Choo, C. A.; Li, Y.; Lu, T. J.; Ittycheriah, A.; Shroff, P.; Sun, P.; Varadarajan, M.; Bahargam, S.; Willoughby, R.; Gaddy, D.; Dasgupta, I.; Desjardins, G.; Cornero, M.; Robenek, B.; Mittal, B.; Albrecht, B.; Shenoy, A.; Moiseev, F.; Jacobsson, H.; Ghaffarkhah, A.; Rivière, M.; Walton, A.; Crepy, C.; Parrish, A.; Liu, Y.

- Zhou, Z.; Farabet, C.; Radebaugh, C.; Srinivasan, P.; van der Salm, C.; Fildjeland, A.; Scellato, S.; Latorre-Chimoto, E.; Klimczak-Plucińska, H.; Bridson, D.; de Cesare, D.; Hudson, T.; Mendolicchio, P.; Walker, L.; Morris, A.; Penchev, I.; Mauger, M.; Guseynov, A.; Reid, A.; Odoom, S.; Loher, L.; Cotruta, V.; Yenugula, M.; Grewe, D.; Petrushkina, A.; Duerig, T.; Sanchez, A.; Yadlowsky, S.; Shen, A.; Globerson, A.; Kurzrok, A.; Webb, L.; Dua, S.; Li, D.; Lahoti, P.; Bhupatiraju, S.; Hurt, D.; Qureshi, H.; Agarwal, A.; Shani, T.; Eyal, M.; Khare, A.; Belle, S. R.; Wang, L.; Tekur, C.; Kale, M. S.; Wei, J.; Sang, R.; Saeta, B.; Liechty, T.; Sun, Y.; Zhao, Y.; Lee, S.; Nayak, P.; Fritz, D.; Vuyyuru, M. R.; Aslanides, J.; Vyas, N.; Wicke, M.; Ma, X.; Bilal, T.; Eltyshiev, E.; Balle, D.; Martin, N.; Cate, H.; Manyika, J.; Amiri, K.; Kim, Y.; Xiong, X.; Kang, K.; Luisier, F.; Tripuraneni, N.; Madras, D.; Guo, M.; Waters, A.; Wang, O.; Ainslie, J.; Baldridge, J.; Zhang, H.; Pruthi, G.; Bauer, J.; Yang, F.; Mansour, R.; Gelman, J.; Xu, Y.; Polovets, G.; Liu, J.; Cai, H.; Chen, W.; Sheng, X.; Xue, E.; Ozair, S.; Yu, A.; Angermueller, C.; Li, X.; Wang, W.; Wiesinger, J.; Koukoumidis, E.; Tian, Y.; Iyer, A.; Gurumurthy, M.; Goldenson, M.; Shah, P.; Blake, M. K.; Yu, H.; Urbanowicz, A.; Palomaki, J.; Fernando, C.; Brooks, K.; Durden, K.; Mehta, H.; Momchev, N.; Rahimtoroghi, E.; Georgaki, M.; Raul, A.; Ruder, S.; Redshaw, M.; Lee, J.; Jalan, K.; Li, D.; Perng, G.; Hechtman, B.; Schuh, P.; Nasr, M.; Chen, M.; Milan, K.; Mikulik, V.; Strohman, T.; Franco, J.; Green, T.; Hassabis, D.; Kavukcuoglu, K.; Dean, J.; and Vinyals, O. 2023. Gemini: A Family of Highly Capable Multimodal Models. ArXiv:2312.11805 [cs].
- Gilbert, T. K.; Lambert, N.; Dean, S.; Zick, T.; Snoswell, A.; and Mehta, S. 2023. Reward Reports for Reinforcement Learning. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '23, 84–130. New York, NY, USA: Association for Computing Machinery.
- Google DeepMind. 2023. AI Safety Summit: An update on our approach to safety and responsibility.
- Guha, N.; Lawrence, C. M.; Gailmard, L. A.; Rodolfa, K. T.; Surani, F.; Bommasani, R.; Raji, I. D.; Cuéllar, M.-F.; Honigsberg, C.; Liang, P.; and Ho, D. E. 2024. AI Regulation Has Its Own Alignment Problem: The Technical and Institutional Feasibility of Disclosure, Registration, Licensing, and Auditing. *The George Washington Law Review*, 92(forthcoming).
- Hadfield, G.; Cuéllar, M. F.; and O'Reilly, T. 2023. It's Time to Create a National Registry for Large AI Models.
- Hendrycks, D.; Carlini, N.; Schulman, J.; and Steinhardt, J. 2022. Unsolved Problems in ML Safety. ArXiv:2109.13916 [cs].
- Hendrycks, D.; Mazeika, M.; and Woodside, T. 2023. An Overview of Catastrophic AI Risks. ArXiv:2306.12001 [cs].
- Kaminski, M. E. 2023. Regulating the Risks of AI. *Boston University Law Review*, 103: 1347–1411.
- Kapoor, S.; Bommasani, R.; Klyman, K.; Longpre, S.; Ramaswami, A.; Cihon, P.; Hopkins, A. K.; Bankston, K.; Biderman, S.; Bogen, M.; Chowdhury, R.; Engler, A.; Henderson, P.; Jernite, Y.; Lazar, S.; Maffulli, S.; Nelson, A.; Pineau, J.; Skowron, A.; Song, D.; Storchan, V.; Zhang, D.; Ho, D. E.; Liang, P.; and Narayanan, A. 2024. Position: On the Societal Impact of Open Foundation Models. In *Forty-First International Conference on Machine Learning*.
- Karkkainen, B. 2008. Bottlenecks and baselines: Tackling information deficits in environmental regulation. *Texas Law Review*, 86: 1409–1444.
- Kolt, N. 2024a. Algorithmic Black Swans. *Washington University Law Review*, 101(1177).
- Kolt, N. 2024b. Governing AI Agents.
- Lambert, N.; Gilbert, T. K.; and Zick, T. 2023. The History and Risks of Reinforcement Learning and Human Feedback. ArXiv:2310.13595 [cs].
- Longpre, S.; Kapoor, S.; Klyman, K.; Ramaswami, A.; Bommasani, R.; Bili-Hamelin, B.; Huang, Y.; Skowron, A.; Yong, Z. X.; Kotha, S.; Zeng, Y.; Shi, W.; Yang, X.; Southen, R.; Robey, A.; Chao, P.; Yang, D.; Jia, R.; Kang, D.; Pentland, A.; Narayanan, A.; Liang, P.; and Henderson, P. 2024. Position: A Safe Harbor for AI Evaluation and Red Teaming. In *Forty-First International Conference on Machine Learning*.
- Madsen, P.; Dillon, R. L.; and Tinsley, C. H. 2015. Airline Safety Improvement Through Experience with Near-Misses: A Cautionary Tale. *Risk Analysis*, 36(5): 1054–1066.
- Manheim, D. 2023. Building a Culture of Safety for AI: Perspectives and Challenges.
- Marchant, G. E.; and Gutierrez, C. I. 2022. Soft Law 2.0: An Agile and Effective Governance Approach for Artificial Intelligence. *Minnesota Journal of Law, Science and Technology*, 24: 375.
- McGregor, S. 2021. Preventing Repeated Real World AI Failures by Cataloging Incidents: The AI Incident Database. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(17): 15458–15463. Number: 17.
- McGregor, S.; Paeth, K.; and Lam, K. 2022. Indexing AI Risks with Incidents, Issues, and Variants. ArXiv:2211.10384 [cs].
- Mitchell, M.; Wu, S.; Zaldivar, A.; Barnes, P.; Vasserman, L.; Hutchinson, B.; Spitzer, E.; Raji, I. D.; and Gebru, T. 2019. Model Cards for Model Reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 220–229. Atlanta GA USA: ACM. ISBN 978-1-4503-6125-5.
- Model Evaluation and Threat Research. 2023. Responsible Scaling Policies (RSPs).
- Mökander, J.; Schuett, J.; Kirk, H. R.; and Floridi, L. 2023. Auditing large language models: a three-layered approach. *AI and Ethics*.
- National Institute of Standards and Technology. 2021. AI Risk Management Framework. *NIST*.
- Norval, C.; Cornelius, K.; Cobbe, J.; and Singh, J. 2022. Disclosure by Design: Designing information disclosures to support meaningful transparency and accountability. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, 679–690. New

York, NY, USA: Association for Computing Machinery. ISBN 978-1-4503-9352-2.

OECD.AI Network of Experts. 2024. Expert Group on AI Incidents.

Ojewale, V.; Steed, R.; Vecchione, B.; Birhane, A.; and Raji, I. D. 2024. Towards AI Accountability Infrastructure: Gaps and Opportunities in AI Audit Tooling. ArXiv:2402.17861 [cs].

OpenAI. 2023a. GPT-4 System Card. Technical report, OpenAI.

OpenAI. 2023b. GPT-4V(ision) system card. Technical report, OpenAI.

OpenAI. 2023c. Preparedness. Technical report, OpenAI.

Phuong, M.; Aitchison, M.; Catt, E.; Cogan, S.; Kaskasoli, A.; Krakovna, V.; Lindner, D.; Rahtz, M.; Assael, Y.; Hodgkinson, S.; Howard, H.; Lieberum, T.; Kumar, R.; Raad, M. A.; Webson, A.; Ho, L.; Lin, S.; Farquhar, S.; Hutter, M.; Deletang, G.; Ruoss, A.; El-Sayed, S.; Brown, S.; Dragan, A.; Shah, R.; Dafoe, A.; and Shevlane, T. 2024. Evaluating Frontier Models for Dangerous Capabilities. ArXiv:2403.13793 [cs].

Raji, I. D.; Xu, P.; Honigsberg, C.; and Ho, D. 2022. Outsider Oversight: Designing a Third Party Audit Ecosystem for AI Governance. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '22, 557–571. New York, NY, USA: Association for Computing Machinery. ISBN 978-1-4503-9247-1.

Reid, M.; Savinov, N.; Teplyashin, D.; Lepikhin, D.; Lillcrap, T.; Alayrac, J.-b.; Soricute, R.; Lazaridou, A.; Firat, O.; Schrittwieser, J.; Antonoglou, I.; Anil, R.; Borgeaud, S.; Dai, A.; Millican, K.; Dyer, E.; Glaese, M.; Sottiaux, T.; Lee, B.; Viola, F.; Reynolds, M.; Xu, Y.; Molloy, J.; Chen, J.; Isard, M.; Barham, P.; Hennigan, T.; McIlroy, R.; Johnson, M.; Schalkwyk, J.; Collins, E.; Rutherford, E.; Moreira, E.; Ayoub, K.; Goel, M.; Meyer, C.; Thornton, G.; Yang, Z.; Michalewski, H.; Abbas, Z.; Schucher, N.; Anand, A.; Ives, R.; Keeling, J.; Lenc, K.; Haykal, S.; Shakeri, S.; Shyam, P.; Chowdhery, A.; Ring, R.; Spencer, S.; Sezener, E.; Vilnis, L.; Chang, O.; Morioka, N.; Tucker, G.; Zheng, C.; Woodman, O.; Attaluri, N.; Kocisky, T.; Eltyshev, E.; Chen, X.; Chung, T.; Selo, V.; Brahma, S.; Georgiev, P.; Slone, A.; Zhu, Z.; Lottes, J.; Qiao, S.; Caine, B.; Riedel, S.; Tomala, A.; Chadwick, M.; Love, J.; Choy, P.; Mittal, S.; Houlsby, N.; Tang, Y.; Lamm, M.; Bai, L.; Zhang, Q.; He, L.; Cheng, Y.; Humphreys, P.; Li, Y.; Brin, S.; Cas-sirer, A.; Miao, Y.; Zilka, L.; Tobin, T.; Xu, K.; Proleev, L.; Sohn, D.; Magni, A.; Hendricks, L. A.; Gao, I.; Ontañón, S.; Bunyan, O.; Byrd, N.; Sharma, A.; Zhang, B.; Pinto, M.; Sinha, R.; Mehta, H.; Jia, D.; Caelles, S.; Webson, A.; Morris, A.; Roelofs, B.; Ding, Y.; Strudel, R.; Xiong, X.; Ritter, M.; Dehghani, M.; Chaabouni, R.; Karmarkar, A.; Lai, G.; Mentzer, F.; Xu, B.; Li, Y.; Zhang, Y.; Paine, T. L.; Goldin, A.; Neyshabur, B.; Baumli, K.; Levskaya, A.; Laskin, M.; Jia, W.; Rae, J. W.; Xiao, K.; He, A.; Giordano, S.; Yagati, L.; Lespiau, J.-B.; Natsev, P.; Ganapathy, S.; Liu, F.; Martins, D.; Chen, N.; Xu, Y.; Barnes, M.; May, R.; Vezer, A.; Oh, J.; Franko, K.; Bridgers, S.; Zhao, R.; Wu, B.; Mustafa,

B.; Sechrist, S.; Parisotto, E.; Pillai, T. S.; Larkin, C.; Gu, C.; Sorokin, C.; Krikun, M.; Guseynov, A.; Landon, J.; Datta, R.; Pritzel, A.; Thacker, P.; Yang, F.; Hui, K.; Hauth, A.; Yeh, C.-K.; Barker, D.; Mao-Jones, J.; Austin, S.; Sheahan, H.; Schuh, P.; Svensson, J.; Jain, R.; Ramasesh, V.; Briukhov, A.; Chung, D.-W.; von Glehn, T.; Butterfield, C.; Jhakra, P.; Wiethoff, M.; Frye, J.; Grimstad, J.; Changpinyo, B.; Lan, C. L.; Bortsova, A.; Wu, Y.; Voigtlaender, P.; Sainath, T.; Smith, C.; Hawkins, W.; Cao, K.; Besley, J.; Srinivasan, S.; Omernick, M.; Gaffney, C.; Surita, G.; Burnell, R.; Damoc, B.; Ahn, J.; Brock, A.; Pajarskas, M.; Petrushkina, A.; Noury, S.; Blanco, L.; Swersky, K.; Ahuja, A.; Avrahami, T.; Misra, V.; de Liedekerke, R.; Iinuma, M.; Polozov, A.; York, S.; Driessche, G. v. d.; Michel, P.; Chiu, J.; Blevins, R.; Gleicher, Z.; Recasens, A.; Rustemi, A.; Gribovskaya, E.; Roy, A.; Gworek, W.; Arnold, S.; Lee, L.; Lee-Thorp, J.; Maggioni, M.; Piqueras, E.; Badola, K.; Vikram, S.; Gonzalez, L.; Baddepudi, A.; Senter, E.; Devlin, J.; Qin, J.; Azzam, M.; Trebacz, M.; Polacek, M.; Krishnakumar, K.; Chang, S.-y.; Tung, M.; Penchev, I.; Joshi, R.; Olszewska, K.; Muir, C.; Wirth, M.; Hartman, A. J.; Newlan, J.; Kashem, S.; Bolina, V.; Dabir, E.; van Amersfoort, J.; Ahmed, Z.; Cobon-Kerr, J.; Kamath, A.; Hrafnkelsson, A. M.; Hou, L.; Mackinnon, I.; Frechette, A.; Noland, E.; Si, X.; Taropa, E.; Li, D.; Crone, P.; Gulati, A.; Cevey, S.; Adler, J.; Ma, A.; Silver, D.; Tokumine, S.; Powell, R.; Lee, S.; Chang, M.; Hassan, S.; Mincu, D.; Yang, A.; Levine, N.; Brennan, J.; Wang, M.; Hodgkinson, S.; Zhao, J.; Lipschultz, J.; Pope, A.; Chang, M. B.; Li, C.; Shafey, L. E.; Paganini, M.; Douglas, S.; Bohnet, B.; Pardo, F.; Odoom, S.; Rosca, M.; Santos, C. N. d.; Soparkar, K.; Guez, A.; Hudson, T.; Hansen, S.; Asawaroengchai, C.; Addanki, R.; Yu, T.; Stokowiec, W.; Khan, M.; Gilmer, J.; Lee, J.; Bostock, C. G.; Rong, K.; Caton, J.; Pejman, P.; Pavetic, F.; Brown, G.; Sharma, V.; Lučić, M.; Samuel, R.; Djolonga, J.; Mandhane, A.; Sjöstrand, L. L.; Buchatskaya, E.; White, E.; Clay, N.; Jiang, J.; Lim, H.; Hemsley, R.; Labanowski, J.; De Cao, N.; Steiner, D.; Hashemi, S. H.; Austin, J.; Gergely, A.; Blyth, T.; Stanton, J.; Shivakumar, K.; Siddhant, A.; Andreassen, A.; Araya, C.; Sethi, N.; Shivan, R.; Hand, S.; Bapna, A.; Khodaei, A.; Miech, A.; Tanzer, G.; Swing, A.; Thakoor, S.; Pan, Z.; Nado, Z.; Winkler, S.; Yu, D.; Saleh, M.; Maggiore, L.; Barr, I.; Giang, M.; Kagohara, T.; Danihelka, I.; Marathe, A.; Feinberg, V.; Elhawaty, M.; Ghelani, N.; Horgan, D.; Miller, H.; Walker, L.; Tanburn, R.; Tariq, M.; Shrivastava, D.; Xia, F.; Chiu, C.-C.; Ashwood, Z.; Baatarsukh, K.; Samangoeei, S.; Al-cober, F.; Stjerngren, A.; Komarek, P.; Tsihlias, K.; Boral, A.; Comanescu, R.; Chen, J.; Liu, R.; Bloxwich, D.; Chen, C.; Sun, Y.; Feng, F.; Mauger, M.; Dotiwalla, X.; Hellendoorn, V.; Sharman, M.; Zheng, I.; Haridasan, K.; Barth-Marón, G.; Swanson, C.; Rogozińska, D.; Andreev, A.; Rubenstein, P. K.; Sang, R.; Hurt, D.; Elsayed, G.; Wang, R.; Lacey, D.; Ilić, A.; Zhao, Y.; Aroyo, L.; Iwuanyanwu, C.; Nikolaev, V.; Lakshminarayanan, B.; Jazayeri, S.; Kaufman, R. L.; Varadarajan, M.; Tekur, C.; Fritz, D.; Khalman, M.; Reitter, D.; Dasgupta, K.; Sarcar, S.; Ornduff, T.; Snider, J.; Huot, F.; Jia, J.; Kemp, R.; Trdin, N.; Vijayakumar, A.; Kim, L.; Angermueller, C.; Lao, L.; Liu, T.; Zhang, H.; Engel, D.; Greene, S.; White, A.; Austin, J.; Taylor, L.; Ashraf, S.;

- Liu, D.; Georgaki, M.; Cai, I.; Kulizhskaya, Y.; Goenka, S.; Saeta, B.; Vodrahalli, K.; Frank, C.; de Cesare, D.; Robenek, B.; Richardson, H.; Alnahlawi, M.; Yew, C.; Ponnappalli, P.; Tagliasacchi, M.; Korchemniy, A.; Kim, Y.; Li, D.; Rosgen, B.; Ashwood, Z.; Levin, K.; Wiesner, J.; Banzal, P.; Srinivasan, P.; Yu, H.; Ünlü, Ç.; Reid, D.; Tung, Z.; Finkelstein, D.; Kumar, R.; Elisseff, A.; Huang, J.; Zhang, M.; Zhu, R.; Aguilar, R.; Giménez, M.; Xia, J.; Dousse, O.; Gierke, W.; Yeganeh, S. H.; Yates, D.; Jalan, K.; Li, L.; Latorre-Chimoto, E.; Nguyen, D. D.; Durden, K.; Kallakuri, P.; Liu, Y.; Johnson, M.; Tsai, T.; Talbert, A.; Liu, J.; Neitz, A.; Elkind, C.; Selvi, M.; Jasarevic, M.; Soares, L. B.; Cui, A.; Wang, P.; Wang, A. W.; Ye, X.; Kallarackal, K.; Loher, L.; Lam, H.; Broder, J.; Holtmann-Rice, D.; Martin, N.; Ramadhana, B.; Toyama, D.; Shukla, M.; Basu, S.; Mohan, A.; Fernando, N.; Fiedel, N.; Paterson, K.; Li, H.; Garg, A.; Park, J.; Choi, D.; Wu, D.; Singh, S.; Zhang, Z.; Globerson, A.; Yu, L.; Carpenter, J.; Quitry, F. d. C.; Radebaugh, C.; Lin, C.-C.; Tudor, A.; Shroff, P.; Garmon, D.; Du, D.; Vats, N.; Lu, H.; Iqbal, S.; Yakubovich, A.; Tripuraneni, N.; Manyika, J.; Qureshi, H.; Hua, N.; Ngani, C.; Raad, M. A.; Forbes, H.; Bulanova, A.; Stanway, J.; Sundararajan, M.; Ungureanu, V.; Bishop, C.; Li, Y.; Venkatraman, B.; Li, B.; Thornton, C.; Scellato, S.; Gupta, N.; Wang, Y.; Tenney, I.; Wu, X.; Shenoy, A.; Carvajal, G.; Wright, D. G.; Bariach, B.; Xiao, Z.; Hawkins, P.; Dalmia, S.; Farabet, C.; Valenzuela, P.; Yuan, Q.; Welty, C.; Agarwal, A.; Chen, M.; Kim, W.; Hulse, B.; Dukkipati, N.; Paszke, A.; Bolt, A.; Davoodi, E.; Choo, K.; Beattie, J.; Prendki, J.; Vashisht, H.; Santamaria-Fernandez, R.; Cobo, L. C.; Wilkiewicz, J.; Madras, D.; Elqursh, A.; Uy, G.; Ramirez, K.; Harvey, M.; Liechty, T.; Zen, H.; Seibert, J.; Hu, C. H.; Elhawaty, M.; Khorlin, A.; Le, M.; Aharoni, A.; Li, M.; Wang, L.; Kumar, S.; Lince, A.; Casagrande, N.; Hoover, J.; Badawy, D. E.; Soergel, D.; Vnukov, D.; Miecznikowski, M.; Simsa, J.; Koop, A.; Kumar, P.; Sellam, T.; Vlasic, D.; Daruki, S.; Shabat, N.; Zhang, J.; Su, G.; Zhang, J.; Liu, J.; Sun, Y.; Palmer, E.; Ghaffarkhah, A.; Xiong, X.; Cotruta, V.; Fink, M.; Dixon, L.; Sreevatsa, A.; Goedeckemeyer, A.; Dimitriev, A.; Jafari, M.; Crocker, R.; FitzGerald, N.; Kumar, A.; Ghemawat, S.; Philips, I.; Liu, F.; Liang, Y.; Sterneck, R.; Repina, A.; Wu, M.; Knight, L.; Georgiev, M.; Lee, H.; Askham, H.; Chakladar, A.; Louis, A.; Crous, C.; Cate, H.; Petrova, D.; Quinn, M.; Owusu-Afriyie, D.; Singhal, A.; Wei, N.; Kim, S.; Vincent, D.; Nasr, M.; Choquette-Choo, C. A.; Tojo, R.; Lu, S.; Casas, D. d. L.; Cheng, Y.; Bolukbasi, T.; Lee, K.; Fatehi, S.; Ananthanarayanan, R.; Patel, M.; Kaed, C.; Li, J.; Sygnowski, J.; Belle, S. R.; Chen, Z.; Konzelmann, J.; Pöder, S.; Garg, R.; Koverkathu, V.; Brown, A.; Dyer, C.; Liu, R.; Nova, A.; Xu, J.; Petrov, S.; Hassabis, D.; Kavukcuoglu, K.; Dean, J.; and Vinyals, O. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. ArXiv:2403.05530 [cs].
- Sastry, G.; Heim, L.; Belfield, H.; Anderljung, M.; Brundage, M.; Hazell, J.; O’Keefe, C.; Hadfield, G. K.; Ngo, R.; Pilz, K.; Gor, G.; Bluemke, E.; Shoker, S.; Egan, J.; Trager, R. F.; Avin, S.; Weller, A.; Bengio, Y.; and Coyle, D. 2024. Computing Power and the Governance of Artificial Intelligence. ArXiv:2402.08797 [cs].
- Schuett, J.; Dreksler, N.; Anderljung, M.; McCaffary, D.; Heim, L.; Bluemke, E.; and Garfinkel, B. 2023. Towards best practices in AGI safety and governance: A survey of expert opinion. ArXiv:2305.07153 [cs].
- Secretary of State for Science, Innovation and Technology. 2023. Policy paper: Introducing the AI Safety Institute. Technical Report E03012924, GOV.UK. ISBN: 978-1-5286-4538-6.
- Shevlane, T.; Farquhar, S.; Garfinkel, B.; Phuong, M.; Whittlestone, J.; Leung, J.; Kokotajlo, D.; Marchal, N.; Anderljung, M.; Kolt, N.; Ho, L.; Siddarth, D.; Avin, S.; Hawkins, W.; Kim, B.; Gabriel, I.; Bolina, V.; Clark, J.; Bengio, Y.; Christiano, P.; and Dafoe, A. 2023. Model evaluation for extreme risks. ArXiv:2305.15324 [cs].
- Shrishak, K. 2023. How to deal with an AI near-miss: Look to the skies. *Bulletin of the Atomic Scientists*, 79(3): 166–169.
- Stephenson, M. C. 2011. Information Acquisition and Institutional Design. *Harvard Law Review*, 124(6): 1422–1483.
- The White House. 2022. Blueprint for an AI Bill of Rights | OSTP.
- The White House. 2023a. Biden-Harris Administration Launches Artificial Intelligence Cyber Challenge to Protect America’s Critical Software.
- The White House. 2023b. Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence. Technical report, The White House.
- The White House. 2023c. FACT SHEET: Biden-Harris Administration Announces New Actions to Promote Responsible AI Innovation that Protects Americans’ Rights and Safety.
- The White House. 2023d. FACT SHEET: Biden-Harris Administration Secures Voluntary Commitments from Leading Artificial Intelligence Companies to Manage the Risks Posed by AI.
- Toner, H.; Ji, J.; Bansemmer, J.; and Lim, L. 2023. Skating to Where the Puck Is Going. Technical report, CSET.
- Touvron, H.; Martin, L.; Stone, K.; Albert, P.; Almahairi, A.; Babaei, Y.; Bashlykov, N.; Batra, S.; Bhargava, P.; Bhosale, S.; Bikel, D.; Blecher, L.; Ferrer, C. C.; Chen, M.; Cucurull, G.; Esiobu, D.; Fernandes, J.; Fu, J.; Fu, W.; Fuller, B.; Gao, C.; Goswami, V.; Goyal, N.; Hartshorn, A.; Hosseini, S.; Hou, R.; Inan, H.; Kardas, M.; Kerkez, V.; Khabsa, M.; Kloumann, I.; Korenev, A.; Koura, P. S.; Lachaux, M.-A.; Lavril, T.; Lee, J.; Liskovich, D.; Lu, Y.; Mao, Y.; Martinet, X.; Mihaylov, T.; Mishra, P.; Molybog, I.; Nie, Y.; Poulton, A.; Reizenstein, J.; Rungta, R.; Saladi, K.; Schelten, A.; Silva, R.; Smith, E. M.; Subramanian, R.; Tan, X. E.; Tang, B.; Taylor, R.; Williams, A.; Kuan, J. X.; Xu, P.; Yan, Z.; Zarov, I.; Zhang, Y.; Fan, A.; Kambadur, M.; Narang, S.; Rodriguez, A.; Stojnic, R.; Edunov, S.; and Scialom, T. 2023. Llama 2: Open Foundation and Fine-Tuned Chat Models. ArXiv:2307.09288 [cs].
- Turri, V.; and Dzombak, R. 2023. Why We Need to Know More: Exploring the State of AI Incident Documentation

Practices. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '23, 576–583. New York, NY, USA: Association for Computing Machinery.

U.S. Department of Commerce. 2023. At the Direction of President Biden, Department of Commerce to Establish U.S. Artificial Intelligence Safety Institute to Lead Efforts on AI Safety.

U.S. Department of Commerce. 2024. Biden-Harris Administration Announces First-Ever Consortium Dedicated to AI Safety.

Van Loo, R. 2019. Regulatory Monitors: Policing Firms in the Compliance Era. *Columbia Law Review*, 119.

Wei, J.; Tay, Y.; Bommasani, R.; Raffel, C.; Zoph, B.; Borgeaud, S.; Yogatama, D.; Bosma, M.; Zhou, D.; Metzler, D.; Chi, E. H.; Hashimoto, T.; Vinyals, O.; Liang, P.; Dean, J.; and Fedus, W. 2022. Emergent Abilities of Large Language Models. ArXiv:2206.07682 [cs].

Weidinger, L.; Barnhart, J.; Brennan, J.; Butterfield, C.; Young, S.; Hawkins, W.; Hendricks, L. A.; Comanescu, R.; Chang, O.; Rodriguez, M.; Beroshi, J.; Bloxwich, D.; Proleev, L.; Chen, J.; Farquhar, S.; Ho, L.; Gabriel, I.; Dafoe, A.; and Isaac, W. 2024. Holistic Safety and Responsibility Evaluations of Advanced AI Models. arXiv:2404.14068.

Weidinger, L.; Rauh, M.; Marchal, N.; Manzini, A.; Hendricks, L. A.; Mateos-Garcia, J.; Bergman, S.; Kay, J.; Griffin, C.; Bariach, B.; Gabriel, I.; Rieser, V.; and Isaac, W. 2023. Sociotechnical Safety Evaluation of Generative AI Systems. ArXiv:2310.11986 [cs].

Whittlestone, J.; and Clark, J. 2021. Why and How Governments Should Monitor AI Development. ArXiv:2108.12427 [cs].

Woodside, T. 2024. Keeping Up with the Frontier: Why Congress Should Codify Reporting Requirements For Advanced AI Systems. Technical report, CSET.