Foundation Model Transparency Reports

Rishi Bommasani[1], Kevin Klyman[1], Shayne Longpre[2], Betty Xiong[1], Sayash Kapoor[3], Nestor Maslej[1], Arvind Narayanan[3], Percy Liang[1]

[1]Stanford University
[2]Massachusetts Institute of Technology
[3]Princeton University
nlprishi@stanford.edu

Abstract

Foundation models are critical digital technologies with sweeping societal impact that necessitates transparency. To codify how foundation model developers should provide transparency about the development and deployment of their models, we propose Foundation Model Transparency Reports, drawing upon the transparency reporting practices in social media. While external documentation of societal harms prompted social media transparency reports, our objective is to institutionalize transparency reporting for foundation models while the industry is still nascent. To design our reports, we identify 6 design principles given the successes and shortcomings of social media transparency reporting. To further schematize our reports, we draw upon the 100 transparency indicators from the Foundation Model Transparency Index. Given these indicators, we measure the extent to which they overlap with the transparency requirements included in six prominent government policies (e.g. the EU AI Act, the US Executive Order on Safe, Secure, and Trustworthy AI). Well-designed transparency reports could reduce compliance costs, in part due to overlapping regulatory requirements across different jurisdictions. We encourage foundation model developers to regularly publish transparency reports, building upon recommendations from the G7 and the White House.

Introduction

Foundation models are transformative digital technologies (Bommasani et al. 2021), introducing new capabilities (Wei et al. 2022) and risks (Weidinger et al. 2022) that bring unprecedented public attention to AI. The potential for pervasive societal impact necessitates greater transparency for foundation models. The 2023 Foundation Model Transparency Index (Bommasani et al. 2023a) confirms that, currently, the foundation model ecosystem is opaque: the Index scored 10 major foundation model developers (e.g. OpenAI, Google, Meta) on a 100 point scale for transparency, with developers on average receiving a mere 37 out of 100.

Previous digital technologies, especially social media platforms, have been similarly plagued by insufficient transparency. Over the past 15 years, social media platforms have come to produce *transparency reports*: public reports, produced on recurring basis, that consolidate information related to usage of their platforms and key platform governance practices like takedown requests and policy enforcement. Today, transparency reports are an industry standard: Access Now documents that more than 85 Internet and telecommunications companies have produced transparency reports (Access Now 2023). The European Union's Digital Services Act mandates transparency reporting for online platforms and formalizes the process to ensure that vital information is publicly reported with sufficient fidelity, frequency, standardization, and accessibility (European Commission 2023).

While transparency practices are nascent for foundation models, and the current landscape displays both idiosyncratic and systematic opacity (Bommasani et al. 2023a), governments are stepping in to take corrective measures. In the United States, Representatives Donald Beyer and Anna Eshoo have introduced the AI Foundation Model Transparency Act to mandate public reporting of standardized information as to be determined by the Federal Trade Commission. This bill builds on reporting requirements from the October 2023 Executive Order on AI. In the European Union, the EU AI Act requires transparency on training data, energy usage, model evaluations, and risk management. And several other jurisdictions such as China, the United Kingdom, and Canada are also taking steps increase transparency.

To address the transparency deficits in the foundation model ecosystem, build upon transparency practices for social media platforms, and guide the transparency initiatives proposed by governments, we propose *Foundation Model Transparency Reports*. Foundation Model Transparency Reports would standardize what companies should report, consolidate this information to assist stakeholders in finding it, and structure the information to facilitate subsequent analysis or comparison across multiple developers. Our transparency reports operationalize recommendations from the White House and G7 that direct foundation model developers to publish transparency reports (Executive Order 14110 2023; Group of Seven 2023). To design our reports, we use the 100 transparency indicators defined in the Foundation Model Transparency Index (Bommasani et al. 2023a) that concretize transparency for foundation models across the supply chain. To measure transparency, Bommasani et al. (2023a) scored foundation model developers on their current practices, whereas we operationalize how developers can

implement new reporting practices to improve transparency.

Using our transparency reports as a lens to study transparency requirements in government policies, we find that current transparency requirements are considerably imprecise. Across 6 major policies (e.g. the US Executive Order on AI, the EU AI Act), we find that 57 of the 100 indicators we include in our reports are not covered by any policy. Compared to these policies, our reports are more precise (e.g. specifying not just what information should be reported, but how and to what level of minimum specificity) and they increase coverage by putting forth a more comprehensive view of transparency. In turn, adoption of our transparency reports would better codify information disclosures from companies.

Our paper makes three contributions to advance transparency. First, we explore how transparency reporting is conducted in other industries to derive principles; we use these principles to design Foundation Model Transparency Reports. Second, we align our design with government policies to show how transparency reports could improve compliance and broaden how governments conceptualize transparency. Third, we instantiate our design with examples of Foundation Model Transparency Report entries from different foundation models based on publicly available information, setting a clear example for future reports. Together, our work guides foundation model developers on how to be more transparent and world governments on how to promote transparency through policy.

Social Media Transparency Reports

The rise of social media platforms over the past ten to twenty years provides a natural parallel for foundation models. Namely, a disruptive and powerful emergent technology came to be widely adopted across societies, thereby intermediating important societal functions such as access to information and interpersonal communication. Social media has been associated with several types of risk, some of which resulted in substantial societal harms (e.g. the Cambridge Analytica scandal, the Rohingya genocide in Myanmar). These harms directly relate to the significant opacity of social media platforms (e.g. how is user data shared, how is contented moderated). Therefore, we describe transparency reporting for social media, where it has emerged as a standard practice, to conceptualize transparency reporting for foundation models.

History

In the context of social media and telecommunications, a *transparency report* is a recurring public report of key metrics related to legal information and takedown requests, as well as policy and intellectual property enforcement for large online platforms (Bankston, Schulman, and Woolery 2017; Trust and Safety Professional Association 2023). Telecommunications companies and social media platforms have gradually provided these reports since 2010, in response to public concerns over their handling of privacy, government surveillance, freedom of speech, and misinformation. Initially, these concerns were triggered by disclosures of dissident information to the Chinese government (Schatz 2006), the FBI's use of the Patriot Act for surveillance (American Civil Liberties Union 2010), and Edward Snowden's subsequent disclosures of NSA surveillance practices (Greenwald 2013). Concerns from users and advertisers would later emerge over moderation practices of harmful content (World Federation of Advertisers 2020), spurring greater transparency into platform policy enforcement.

To examine how transparency reports emerged, we consider Google's transparency report in 2010. In 2010, Google first reported government requests for content removal or information (goo 2010), as well as where its services were blocked or inaccessible (Bankston, Schulman, and Woolery 2017). The report showed that Google received over 1200 requests from 36 jurisdictions, providing a greater insight into the platform: for example, Brazil made 398 requests to remove 19000 items, of which Google complied with 68. Shortly thereafter, LinkedIn, Microsoft, and Twitter began producing transparency reports, with an avalanche of uptake following the Snowden revelations in 2013 Access Now's Transparency Reporting Index documents this increase in adoption: 6 companies produced reports in 2012, compared to over 60 in 2015. Transparency reporting also gradually expanded in scope to include removals under intellectual property law and the Digital Millennium Copyright Act (Access Now 2023). Etsy was the first platform, in 2015, to introduce a policy enforcement report, detailing its responses to user violations of its terms of service (ets 2014). Since reporting on these incidents could compromise user privacy, companies generally release high-level aggregate statistics. Overall, transparency reports came to be an important part of companies' brands and helped foster wider public trust and accountability (Bankston, Schulman, and Woolery 2017; Trust and Safety Professional Association 2023).

While social media platforms played a significant role in conceptualizing the first versions of transparency reports, civil society organizations drove advances in their scope and utility. For example, organizations began to rank online platform transparency practices to generate pressure. The Electronic Frontier Foundation regularly scores corporations on privacy, process, and freedom of speech, in "Who Has Your Back" (Crocker et al. 2019). In response, high-scoring companies like WordPress and Apple publicized their results (Zhu 2015). Similarly, Ranking Digital Rights maintains its Corporate Accountability Index, to score telecommunications providers on a spectrum of transparency, access and responsibility to users (MacKinnon et al. 2019). And in response to the Transparency Reporting Toolkit from New America and the Berkman Klein Center (Budish, Woolery, and Bankston 2016), Twitter revamped its reporting standards to follow suggested best practices (Kessel 2016).

Subject to the recommendations of civil society, and the associated push for greater accountability, transparency reporting from major social media platforms had evolved to become more interpretable to the general public, more detailed, and more regular in its cadence. By 2021, 88 technology companies had published transparency reports, with

some including downloadable data (Access Now 2023). For instance, Meta now releases comprehensive and often near-live reports on policy enforcement, intellectual property, government requests, content restrictions, regulatory measures, Internet disruptions, and even widely viewed content (Facebook 2023). In addition, Meta offers content libraries with APIs for Facebook and Instagram, as well as an ad Library. However, since 2021 there has been a steep decline in new voluntary transparency reporting from major platforms (Rydzak 2023); X, for example, no longer makes updates to its Transparency Center (X 2023).

Purpose

In social media, transparency reporting functions as an instrument for social media companies to make information public. In particular, these disclosures help alleviate informational deficits on public interest matters spanning privacy, free speech, surveillance, and the reach of harmful content. Social media platforms are often incentivized to comply with the unethical or secretive requests of governments in order to maintain access to their markets (Gorwa and Ash 2020). While transparency reporting cannot fully deter this incentive, it can inform the public of the scope and extent of a government's intervention into platforms, and spur public pressure as a deterrent to surveillance, censorship, or privacy violations. Additionally, as social media platforms have been likened to a "digital public square," the processes governing the access and dissemination of speech can have significant societal impact (Lazar 2023). In light of rising concerns of algorithmic dissemination, echo chambers, and scalable misinformation, transparency reporting could mediate public trust. In theory, open and transparent processes around speech suppression or amplification would enable a fairer public discourse that is better informed about the measures taken by social media companies to regulate online speech.

Given that the information made available through transparency reports is intrinsically highly multifunctional, transparency reports are simultaneously targeted at a range of stakeholders in the complex platform ecosystem. Nonexhaustively, these stakeholders include platform users, non-users that are impacted by platform operations, investors, and advertisers. Users, non-users, and civil society collectively are invested in ensuring that processes that govern platform information are fair and privacy-preserving. These concerns are partially addressed by clear documentation of standards and procedures for privacy, compliance with governments, and content moderation, as well as public statistics. Consequently, civil society organizations have outlined clear criteria by which transparency reports can better satisfy these stakeholder objectives (Llansó and Vogus 2021; Santa Clara Principles 2023; Aspen Institute 2021). Similarly, it is in advertisers' interests for their ads to not be associated with offensive or harmful content (World Federation of Advertisers 2020). Certain platforms have regulated political advertising, providing a clear example where monitoring the compliance and impact of company policies can provide a useful basis for academic research by social scientists (Edelson et al. 2021). Lastly, in the absence of corporations supporting public interest research, some have argued that society's abilities to understand and address misinformation, among other harms, is severely limited (Abdo et al. 2022).

Implementation

Modern social media transparency reports are typically divided into four categories: legal information requests, legal takedown requests, intellectual property enforcement, and policy enforcement (Trust and Safety Professional Association 2023). Legal information requests typically pertain to requests for private information on users and their communications (Vermeulen 2021). Legal takedown requests pertain to governments applying local laws to have content permanently removed from platforms. Platforms may not always comply with government requests, so reports often show the number of requests by country, the compliance rate, and the number of unique accounts affected. Intellectual property reporting is often split into content removals and requests for copyrighted and trademarked content. Policy enforcement reporting includes a wide range of potential violations, which will differ by platform, and usually display the removal rates over time per country for each violation type. Additionally, platforms may report other metrics that detail the security of user accounts, the content that is most viewed on the platform, or changes in company policies.

While this high-level standardization is common across social media companies, further standardization has been challenging. Primarily, as Keller (2021) outlines, most metrics are not straightforward to calculate and come with implicit assumptions. As a result, given the significant heterogeneity in social media platforms, this requires bespoke, company-specific measurement approaches that inhibit apples-to-apples comparisons (Trust and Safety Professional Association 2023). Because reported statistics are often not standardized, platforms have substantial discretion to select what they measure and how they measure it (Urman and Makhortykh 2023). This idiosyncratic approach intensifies concerns that transparency reporting in social media serves as a type of ethics-washing that is instrumentalized as marketing collateral (Zalnieriute 2021) Further, transparency reporting introduces substantial costs for social media platforms (Stoughton and Rosenzweig 2022). Significant company-internal infrastructure is required to initially measure and subsequently maintain metrics, especially given many social media platforms operate across many jurisdictions. Companies need to dedicate significant resources to maintaining their transparency reports, causing some to question whether this comes at the cost of further investment into more substantive governance or risk mitigation at these organizations.

Serious critiques of transparency reporting—and the broader focus on improving the procedural transparency of digital technology providers—have been raised in various fields (Boyd 2016; Ananny and Crawford 2018; Ghosh and Faxon 2023; Mittelstadt 2019; Han 2015; Birchall 2021). For example, social media companies who release transparency reports rarely sufficient access to their platforms for third parties to validate the information they disclose, meaning the information may be inaccurate. These critiques are

often valid: transparency is not an end unto itself, it is merely a mechanism that may allow further insight into the operations of technology companies in order to better pursue other more tangible societal goals (Bommasani et al. 2023c). The way in which transparency requirements are implemented can have a significant impact on whether transparency is performative and unverifiable or substantive and rigorous.

Mandates

Historically, transparency reporting has been a voluntary practice and, increasingly, an expected norm in the social media industry due to public pressure. However, a growing number of governments are considering mandating transparency reporting. In the United States, other types of disclosure requirements imposed by the government are at times in tension with the First Amendment due to concerns of compelled speech. In ruling on disclosure requirements across several contexts, the Supreme Court has used several different legal standards for distinct types of disclosure, sharing a common basis in requiring the government to prove a disclosure requirement "is appropriately tailored to a sufficiently important goal" (Brannon et al. 2023).

Under the European Union's recently-enacted Digital Services Act (DSA), online platforms are required to abide by transparency and access provisions (Miller 2023). The EU designates Very Large Online Platforms (VLOPs) as platforms with at least 45 million monthly active users (i.e. 10% of the EU population): these platforms must prepare biannual transparency reports, conduct periodic risk assessments, publish audit reports, establish ad repositories, and share data with external researchers. The EU solicited external input on the form and manner of these transparency reports from December 2023 to January 2024, and intends to adopt an implemented regulation in 2024 (European Commission 2023). Primed by the experiences of transparency reporting over the past decade, the EU aims to standardize reporting by identifying a series of indicators that must be reported, along with clarifying measurement methodology in several cases (Schneider, Siegrist, and Oles 2023). In the first round of transparency reporting under the DSA, 19 platforms submitted reports spanning human resources dedicated to content moderation by locale, content enforcement takedown rates and error rates on both content and accounts, as well as the median time needed to enforce content violating the law or platform policy (Commission 2022). This level of specificity has allowed for far greater clarity into operations: for example, the transparency report from X demonstrates glaring disparities in content moderation staffing across languages (e.g. Bulgarian, Croatian, Dutch, Hebrew, Italian, Latvian and Polish all have at most 2 content moderation staff who are primary language speakers, compared to over 2000 for English).[1]

Foundation Model Transparency Reports

To design Foundation Model Transparency Reports, we identify 6 design principles, informed directly by the strengths and weaknesses of social media transparency reporting. Subject to these principles, we then identify indicators to be included in the reports using the Foundation Model Transparency Index (Bommasani et al. 2023a) and work through a few examples of how developers may report information related to these indicators.

Principles

Social media transparency reports, especially in their current form, embody several desirable principles for transparency reporting. First, these reports consolidate information about a social media platform's practice into a *centralized* location, referring both to the transparency report document and the transparency report page on the platform's website. Consolidation and centralization enable stakeholders to have a singular and predictable source for finding relevant information. Second, these reports are *structured* to address specific queries: reports often have four top-level sections. This structure sets clear expectations for what can be found in the report, what the report is unlikely to cover, and as a coarse means for comparing different platform practices. Third, some companies prepare extensive transparency reports that clearly *contextualize* information. Given that transparency reports are read by a variety of stakeholders with differing expertise and familiarity about platforms, and there are many unique nuances of a platform (e.g. what a "user" is in the context of the platform), context is necessary to adequately interpret information.

However, social media transparency reports at present (generally) fail to implement other desirable principles for transparency reporting. First, while these reports consolidate information, the underlying information to be included is not *independently specified*. Consequently, platforms are able to determine what information to include and exclude, allowing them to unevenly report only on matters advantageous to them. Second, while these reports are coarsely structured, they are not fully *standardized* both in terms of the form and organization as well as the indicators reported. Therefore, transparency reports from different platforms cannot be easily compared to each other or combined to perform larger-scale analyses that reveal aggregate trends. Third, while the best transparency reports at present contextualize information, they often do not clearly specify *methodologies* for computing statistics. As a result, given many quantities could be computed in different ways (e.g. different methods of user de-duplication for user counting), without clarity on the underlying methodology, consumers of transparency reports may still be prone to misinterpretation.

Approach

Using these 6 principles—centralization, structure, contextualization, independent specification, standardization, and methodologies—we design Foundation Model Transparency Reports. To begin, rather than having foundation model developers dictate what is included in their own transparency reports, we propose a uniform set of indicators to be included in transparency reports across foundation model developers. This ensures that the contents of the reports are simultaneously (i) independently specified and

[1] https://transparency.twitter.com/dsa-transparency-report.html.

(ii) standardized. To select these indicators, we use the 100 transparency indicators (see Appendix) from the Foundation Model Transparency Index (FMTI; Bommasani et al. 2023a), which is a recent initiative that scores major foundation model developers for their transparency. FMTI provides a comprehensive conceptualization of transparency with its 100 indicators organized into 3 domains: (i) the *upstream* resources used to build a foundation model, (ii) the *model* properties including evaluations, and (iii) the *downstream* use and impact of the foundation model. Domains are further broken down into subdomains (e.g. upstream resources include data, labor, compute, code); we re-use this hierarchical domain-subdomain structure as the recommended organization for Foundation Model Transparency Reports.

In contrast to social media platforms, where platform activities and usage are almost exclusively conducted on the platform's website, foundation models have different usage patterns. As a consequence, transparency reports for foundation models should not only be made available on the foundation model developer's website, but also via distribution channels that make the foundation model available. For example, Meta's Llama 2 model is distributed via Meta's GitHub repository, but also via Microsoft Azure, Hugging Face, and other platforms. As a result, transparency reports would, ideally, be disseminated through these distribution channels as well to ensure the associated information can be discovered even if a consumer of the information does not look for it on Meta's website. Further, akin to the centralization of DSA Transparency Reports in an EU database, governments may consider consolidating Foundation Model Transparency Reports across foundation model developers into a single database to facilitate research and analysis.

Therefore, our design addresses 4 of the 6 principles we identify: independent specification, consolidation/centralization of information, report structure, and standardized information/indicators. In practice, foundation model developers may choose to not be transparent about certain indicators for a variety of reasons such as (i) the costs of generating the relevant information, (ii) the liability risk from disclosing the information, or (iii) the competitive risk from disclosing the information. In these cases, we encourage companies to still include these fields in their transparency reports to make clear to other stakeholders this information is not available and, when possible, to justify why this information is not provided. For example, OpenAI clearly indicates that it is not transparent on several matters (e.g. training data, model size) for GPT-4 as a matter of competition and safety (OpenAI 2023).

To address the final 2 principles of contextualization and (measurement) methodology, we provide three examples that address indicators across the 3 domains (upstream, model, downstream).

Example: Upstream environmental impact. Two of the transparency indicators we include address the direct environmental impact (due to electricity usage) and the broader environmental impact (e.g. due to water used to cool data centers) associated with building the foundation model. As an exemplar of how to provide this transparency, we consider the work of Luccioni, Viguier, and Ligozat (2022) in estimating the environmental impact of training BLOOM (Le Scao et al. 2022) to underscore three matters. This work makes clear what is being reported (i.e. environmental impacts associated with the equipment manufacturing, model training, and model deployments phases) and what assumptions are made (e.g. how different greenhouse gas emissions are converted to tons of carbon dioxide). Beyond this conceptual clarity, the work provides methodological clarity (e.g. in how total emissions are computed as the sum of infrastructure, idle, and dynamic consumption), highlighting components neglected in other environmental accounting approaches (Patterson et al. 2021). Finally, since almost all assessments of environmental impact will hinge on underlying estimates (e.g. the carbon intensity of the energy grid), the reporting is clearly contextualized with the sourcing of this information (in this case to statistics provided by Aurora Energy Research on French carbon utilization).

Example: Model evaluations. Several of the transparency indicators we include address model evaluations that span capabilities, limitations, risks, mitigations, trustworthiness, and efficiency. Unlike the environmental impact example, here we instead describe demonstrated issues and challenges in reporting evaluation results with the standard MMLU (Hendrycks et al. 2021) benchmark for language models. To ensure evaluation results are correctly interpreted, developers should clearly report the resources involved in adapting (e.g. prompting, fine-tuning) their foundation model to the evaluation. For example, Google reports the results for Gemini (Pichai and Hassabis 2023) on MMLU in direct comparison to GPT-4, obscuring that Gemini was prompted using 32 examples and chain-of-thought prompting whereas GPT-4 was prompted using 5 examples and standard in-context learning. Further, developers should specify lower-level details about model evaluations (e.g. the specific prompts used, the codebase and implementation for the evaluation). Fourrier et al. (2023) demonstrates that different implementations of MMLU can lead to noticeably different quantitative results, sometimes even changing the ranking of different models.

Example: Downstream policy enforcement. Several of the transparency indicators correspond with transparency sought for content moderation on social media platforms. Namely, these are indicators on the usage policy for foundation model, the policy's enforcement, the frequency of usage policy violations, the rate of accurate detection of these violations, and whether users are informed about and can appeal moderation decisions. For some of these indicators, producing the relevant information should be of marginal cost to foundation model developers, but we highlight that estimating the rate of usage policy violation is less straightforward. Calculating the prevalence of *total* usage policy violations is more involved than just reporting the number of *detected* usage policy violations, since many usage policy violations may go undetected. To address this issue, Narayanan and Kapoor (2023) provide guidance informed by social media practices. For example, social media companies sample posts uniformly at random to generate estimates of specific policy violations (e.g. hate speech) using human moderation. Foundation model developers could emulate this prac-

tice or use other sampling methods to provide better estimates of total violations and detected violations, identifying potential gaps between the two.

Policy Alignment

Governments are considering or implementing policies to improve transparency. We analyzed 6 major policies (e.g. the EU AI Act, the US Executive Order on AI) to systematically identify correspondences governments' transparency requirements and our transparency reports. The identified correspondences incentivize foundation model developers to report this information (e.g. when it is also required by law) and clarify the differing priorities across jurisdictions. We find fairly limited alignment between these policies and our indicators in some cases, which is often caused by insufficient precision in governments' transparency requirements.

Tracked Policies

We identify 6 major policies (Table 1) that govern foundation models and include transparency obligations. In some cases, these policies explicitly indicate that foundation model developers should prepare transparency reports. The US White House Commitments include a pledge that developers will release transparency reports for foundation models that "include the safety evaluations conducted (including in areas such as dangerous capabilities, to the extent that these are responsible to publicly disclose), significant limitations in performance that have implications for the domains of appropriate use, discussion of the model's effects on societal risks such as fairness and bias, and the results of adversarial testing conducted to evaluate the model's fitness for deployment." The G7 Code of Conduct includes similar provisions for on transparency reports, including transparency regarding human rights and risk evaluation. In both cases, while the White House and the G7 call for transparency reports, they do not clear what all should be contained in the reports, how this information should be reported or who it should be disclosed to.

EU AI Act. The EU AI Act, which was successfully negotiated in December 2023 and will be published in mid 2024, comprehensively regulates AI. Given its broad scope, we highlight the contents relevant for foundation models. The Act creates an AI Office as a new institution with the EU's executive body, the European Commission. The Office will oversee enforcement for general-purpose AI and coordinate with national regulatory authorities across the EU's 27 member states.

US Executive Order on AI. Executive Order 14410 was published in October 2023 by the Biden administration to articulate US AI policy to attract talent, ensure security, and protect civil rights. The order has sweeping scope, introducing over 150 requirements for federal agents to complete, mostly within a year's time (Meinhardt et al. 2023). For foundation models, in addition to reporting requirements for some developers of dual-use foundation models, the Executive Order address topics like large computing clusters, bio-related customer screening, content provenance, and open foundation models.

US AI Foundation Model Transparency Act. The Foundation Model Transparency Act is a legislative proposal introduced by Representatives Anna Eshoo and Don Beyer in December 2023. The bill explicitly targets transparency in the foundation model ecosystem with a particular focus on data both for training and inference. The US Federal Trade Commission would be tasked with developing and enforcing transparency requirements in consultation with the National Institute for Standards and Technology and the Office of Science and Technology Policy.

US White House Voluntary Commitments. The US voluntary commitments are eight commitments made by companies developing AI systems that were announced in two rounds by the White House in July and September 2023.[2] In addition to a commitment to publicly release transparency reports for foundation models, the commitments address red teaming, cybersecurity, watermarking, and bias. The commitments are not retroactive: they "apply only to generative models that are overall more powerful than the current industry frontier" in the case of companies that signed in July, and "they apply only to generative models that are overall more powerful than the current most advanced model produced by the company making the commitment" in the case of companies that signed in September. The US voluntary commitments are intended "to remain in effect until regulations covering substantially the same issues come into force."

Canada Voluntary Code of Conduct. Canada's voluntary code of conduct was announced in September 2023 and has been endorsed by 22 organizations.[3] The code of conduct includes specific measures related to different aspects of responsible development and deployment of foundation models, such as accountability, safety, fairness, human oversight, and robustness. It directs different measures toward developers and managers of generative AI systems, where managers are organizations that put a system into operation, control access, and conduct monitoring (e.g. developers are responsible for mitigating safety risks, while managers must clearly identify AI-generated content). Additionally, the code of conduct distinguishes between obligations of developers and managers of all advanced generative AI systems as opposed to those that are made available for public use. Similar to the US voluntary commitments, "the code identifies measures that should be applied in advance of binding regulation pursuant to the Artificial Intelligence and Data Act by all firms developing or managing the operations of a generative AI system with general-purpose capabilities."

G7 International Code of Conduct. The G7, which includes the US, Canada, and the EU as members, issued its

[2]The July signatories are Amazon, Anthropic, Google, Inflection, Meta, Microsoft, and OpenAI. The September signatories are Adobe, Cohere, IBM, Nvidia, Palantir, Salesforce, Scale AI, and Stability AI.

[3]The signatories are Ada, AlayaCare, Alberta Machine Intelligence Institute, AltaML, Appen, BlackBerry, BlueDot, CGI, Cohere, Council of Canadian Innovators, Coveo, IBM, kama.ai, Mila, OpenText, Protexxa Inc., Ranovus, Resemble AI, Responsible Artificial Intelligence Institute, Scale AI, TELUS, and Vector Institute.

Policy	Status	Type	Covered Entities	Reference
Canada Code of Conduct	In effect	Voluntary	general-purpose generative AI	ISED Canada (2023)
EU AI Act	Negotiated	Mandatory	general purpose AI models	European Council (2024)
G7 Code of Conduct	In effect	Voluntary	the most advanced AI systems	Group of Seven (2023)
US Executive Order	In effect	Mandatory	dual-use foundation models	Executive Order 14110 (2023)
US FM Transparency Act	Proposed	Mandatory	foundation models	United States Congress (2023)
White House Commitments	In effect	Voluntary	AI systems	White House (2023)

Table 1: Government policies with transparency requirements. Information on the 6 policies we examine: policy name, implementation status as of February 1, 2024, the type of transparency requirements, the entities subject to the requirements, and the reference text we analyze.

International Code of Conduct in October 2023 as part of the Hiroshima AI Process. The code of conduct includes provisions on transparency reporting as well as 10 measures related to data protection, risk management, and development of technical standards. While this code of conduct is voluntary and companies have not publicly acceded to it as they have national-level commitments, it may be the basis for a future global agreement.

Measuring Alignment

For each policy, we identify all transparency requirements and tag which indicators in our transparency reports, if any, they correspond to (see Appendix for the 100×6 indicator-policy matrix). Table 2 summarizes the results: on average, the 6 policies share 10 transparency requirements with our 100 transparency indicators. Of the 100 indicators, 43 are covered by at least one policy. That is, there are 57 transparency indicators we include that are not included in any of these policies.

The government transparency requirements rarely focus on the upstream resources required to build foundation models, such as data, labor, and compute. 3 of the 6 policies include no upstream requirements, though the US AI Foundation Model Transparency Act has 10 such requirements, including disclosure of data size, data sources, data augmentation, and personal information in the data. The EU AI Act had the most transparency requirements, including 12 requirements that no other policy contained such as the duration of model development, energy usage, model components, and the model license. The US Executive Order on AI had the fewest transparency requirements of all policies considered and, notably, was the sole policy to require only that companies disclose information to the government.

Common transparency requirements across policies included disclosing data sources (required by 3 policies), centralized model documentation (3), prohibited, restricted, and prohibited uses (3), whether a person is interacting with an AI system (3), documentation for responsible downstream use (3), a capabilities description (4), a risk description (4), evaluation of unintentional harm (4), evaluation of intentional harm (5), a limitations description (5), and a mitigations description (5). These commonalities show shared priorities but also how current transparency requirements are often superficial. In contrast, while our indicators also include descriptions of capabilities, limitations, risks, and mitigations, we further include indicators about the rigorous evaluation of these matters. In the upstream domain, the only consistent transparency requirement relates to data sources, with no policies including transparency requirements related to data labor and only one policy (the AI Act) with a requirement on technical methods.

There were also a handful of transparency requirements that are included in these policies that were not featured in the Foundation Model Transparency Index. For example, Canada's code of conduct requires that developers "maintain a database of reported incidents after deployment, and provide updates as needed to ensure effective mitigation measures." It also requires that firms that manage the operations of generative AI systems "share information and best practices on risk management with firms playing complementary roles in the ecosystem." The EU AI Act contains a number of additional transparency requirements, ranging from the date the model was released to the maximum context window length, the rationale for key design choices, and adverse event reporting. We discuss adverse event reporting, which appears in both of these policies, subsequently.

On the whole, our analysis reveals three core findings. While governments are taking steps to improve transparency, the coverage of their requirements is relatively narrow and not comprehensive: our transparency reports provided a more holistic and organized approach. When consulting these requirements, we find they lack specificity: for example, they do not detail the minimal level of precision required of developers in reporting quantitative information nor do articulate how broader concepts like "risk" should be understood. Given the overlaps we identify, Foundation Model Transparency Reports could systematize how transparency is operationalized, clarify what should be disclosed, and reduce costs by reducing redundancies.

Example of Transparency Report Entries

We provide an example of transparency report entries to demonstrate a Foundation Model Transparency Report. The full report is in the Appendix: we describe how we built it and key takeaways here.

Policy	Transparency for whom	# Upstream	# Model	# Downstream	Total
Canada Code of Conduct	Public, Firms	1	3	5	9
EU AI Act	Public, Firms, Government	9	13	8	30
G7 Code of Conduct	Public	0	7	5	12
US Executive Order	Government	0	4	1	5
US FM Transparency Act	Public	10	7	3	20
US White House Commitments	Public	0	6	1	7

Table 2: Alignment between government policies and our Foundation Model Transparency Reports. For each policy, we indicate which entities receive the disclosed information as well as the overlap between the policy's requirements and our Transparency Report indicators. We report the overlap in aggregate as well as for (i) upstream resources, (ii) model-level properties, and (iii) downstream use. Overall, the transparency requirements in all 6 policies are considerably less comprehensive and less specific than the 100 indicators we consider.

Construction

The 2023 Foundation Model Transparency Index (FMTI) confirms that current transparency practices across the foundation model ecosystem are lackluster. Most major foundation model developers do not provide information on over half of our 100 indicators (Bommasani et al. 2023a). As a result, for the purposes of building an illustrative sample report, we provide examples of transparency report entries from 9 foundation model developers instead of reporting on the practices of a single developer. While in practice, a transparency report will correspond to the practices of a single developer (and be associated with a single model or model family), we nonetheless believe this amalgam serves a useful demonstration.

To create these examples, we consider the 10 foundation model developers scored in FMTI and the associated scores. For any indicator (82 of the 100) where at least one developer scores the point, we consider all developers that receive a point for that indicator. Of these developers, we select one of the developers whose practices best exemplify transparency for the indicator, with some consideration for selecting different developers to portray a variety of practices. Given the selected (indicator, developer) pairs, we then prepared the example entry in the Appendix that states what the developer discloses for the indicator (e.g. Meta's disclosure of development duration for Llama 2, Inflection's description of the limitations of Inflection-1). While this report reflects some of the best existing practices for each indicator, we note that this should not be seen as the ceiling for transparency in many cases.

Analysis

Our examples of Foundation Model Transparency Report entries, in line with findings of the FMTI, implicitly denotes 18 indicators where no major developer is currently transparent (e.g. several labor-related indicators, several indicators on usage statistics and impact). Consequently, developers or others in the community that demonstrate how information regarding these 18 indicators should be disclosed would establish a meaningful precedent. Further, even for many of the 82 indicators where the report contains an entry, significantly more could be done to make this information useful and actionable. In many cases, the level of contextualization and methodological clarity could be specifically improved (e.g. regarding the aspects of model development that contributed to Meta's measurement of duration, or specific limitations were identified by Inflection?).

At a more fine-grained level, certain indicators also reveal how foundation model developers conceptualize model development differently, and the community lacks a common conceptual framework. For example, while practices from Anthropic, Hugging Face/BigScience, Meta, and OpenAI are all included in the example report on the matters of data and labor, different developers describe their data pipelines in substantively different ways. In turn, articulating where human labor is involved and what data processing occurs through this pipeline may yield inconsistent answers across developers that may not be directly comparable. For many indicators in the report, it is unclear if the disclosed information is a partial or complete answer. For example, while policies from Anthropic, Google, Inflection and OpenAI are all included on the matters of terms of service and usage policy enforcement, in several cases it remains unclear whether they capture an exhaustive list of violative behavior, enforcement actions, and associated appeals/justifications. In fact, in some cases this information was only identified by triggering detected usage policy violations by Bommasani et al. (2023a), which brings into question the extent to which these usage policies are fully transparent at present.

Related Work

Transparency is a fundamental value with a significant history of study in AI (Gebru et al. 2021; Bender and Friedman 2018; Mitchell et al. 2018; Raji and Buolamwini 2019; Gray and Suri 2019; Crawford 2021; Vogus and Llansó 2021; Keller 2022; Bommasani et al. 2023b). Here we consider how our approach relates to other transparency methodologies in AI (namely model cards, data sheets, and ecosystem cards) and other reporting methodologies in society (namely financial reporting and adverse event reporting).

While Foundation Model Transparency Reports draw greatest inspiration from social media transparency reports, these other methodological approaches to transparency and reporting can help inform and complement more comprehensive transparency reporting.

Transparency Approaches in AI

The most common approach for improving in transparency in AI is model *evaluations*: these evaluations help to clarify model strengths and weaknesses, often for technical AI practitioners (Bommasani et al. 2023c). While evaluations can provide significant insight into a specific model, they are still limited in their ability to account for broader societal context (e.g. data, labor, downstream impact). In turn, *documentation*-based approaches to increasing transparency play a complementary role to evaluations, often providing legibility to stakeholders beyond technical AI practitioners. While model evaluations characterize a specific model in isolation, documentation situates model and system development in a broader context.

Documentation in AI was pioneered by data sheets (Gebru et al. 2018) and model cards (Mitchell et al. 2018) for data and models, respectively. These documentation frameworks enumerate a series of questions that an AI developer should answer, which are often fairly open-ended and unstructured in form. For example, Gebru et al. (2018) introduces three questions on the motivation for dataset creation: what was the purpose for dataset creation, who created the dataset (and, potentially on whose behalf), and who funded the dataset? These documentation approaches tend to be very comprehensive in their maximal instantiation, which means empirically there is significant heterogeneity in how different organizations produce data sheets or model cards, including which of the original questions posed by Gebru et al. (2018) and Mitchell et al. (2018) are (satisfactorily) addressed. For example, the Llama 2 model card contains most of the high-level categories specified in the original model cards paper, but several of the lower-level questions posed in the paper are not addressed. Another important example of documentation in AI are the reproducibility checklists required by conferences like NeurIPS and EMNLP, which are mandatory for all papers and include various transparency requirements related to training, licensing, and limitations.

More recently, Bommasani et al. (2023b) introduced ecosystem cards as a documentation framework, which akin to our work specifically targets the foundation model setting. Three variants of the ecosystem card template exist for documenting datasets, foundation models, and applications/products respectively, with Bommasani et al. (2023b) emphasizing the importance of tracking dependency relationships between these different assets. In contrast to data sheets and model cards, which were principally envisioned as developer-driven forms of transparency, ecosystem cards can be created and maintained by other actors in the ecosystem.

Relative to these documentation frameworks, Foundation Model Transparency Reports share common themes of organizing information and, in several instances, specific indicators. However, our transparency reports adopt the more comprehensive view of transparency put forth in the Foundation Model Transparency Index, spanning elements across the supply chain. Further, our transparency reports are closer in style to social media transparency reports, with a greater emphasis on more targeted informational queries rather than more open-ended questions found in data sheets and model cards. Our focus is on transparency that is relevant for public accountability and risk management in relation to (widely-deployed) foundation models, whereas many of these prior frameworks are aimed at AI researchers to promote better scientific practices.

Reporting Approaches in Society

In mature industries, companies and organizations are often required to produce reports that document their operations (e.g. tax reporting, environmental reporting, product safety reporting). We consider US financial reporting as a horizontal practice spanning industries, as well as the US Food and Drug Administration's (FDA) adverse event reporting system as a domain-specific practice. These reporting approaches, along with social media transparency reporting, provide additional references in envisioning, designing, and implementing transparency reporting for foundation models.

Financial reporting. In the United States, several overlapping reporting mechanisms provide transparency on the financial ecosystems. Financial reporting is overseen by the Securities and Exchange Commission (SEC), whose mandate is to inform and protect investors, regulate securities markets, and enforce federal securities law (Secutiries and Commision 2024a). The SEC requires that publicly traded companies release significant information about their finances through annual reports (Form 10-K), quarterly reports (Form 10-Q), and current reports (Form 8-K) (Secutiries and Commision 2024c). The 10-K and 10-Q comprehensively characterize a company's financial health (e.g. information on business activities, risk factors, assets, liabilities) (Secutiries and Commision 2024b), whereas the 8-K is required to notify the SEC, and later the public, of sudden events such as bankruptcy or acquisition of significant assets (Secutiries and Commision 2024c). Standards also heavily influence financial reporting. The Generally Accepted Accounting Principles determine accounting standards accepted by the SEC and function as the default for American companies (Secutiries and Commision 2024d), with the International Financial Reporting Standards functioning as their international counterpart. Additionally, the non-profit Public Company Accounting Oversight Board (PCAOB) develops auditing standards for public companies and SEC-registered brokers and dealers. These standards are important as they ensure that business audits are standardized, high quality, and trustworthy (Secutiries and Commision 2024e).

The history of US financial regulation has several instructive lessons for transparency reporting for foundation models. Many American financial reporting mechanisms came out of regulatory measures intended to address issues of low transparency and their subsequent negative effects. The SEC was created in 1934 by Franklin Delano Roosevelt in the aftermath of the 1932 Pechora Commission, which highlighted how abusive practices in the financial industry con-

tributed to the 1929 stock market crash (Perino 2010). Likewise, the PCAOB was created as part of the 2002 Sarbanes-Oxley Act, which was itself a response to major corporate accounting scandals like Enron, WorldCom and Tyco. Beyond creating the PCAOB, the Sarbanes-Oxley Act instrumentally reformed corporate governance and financial disclosure practices in the US, mandating greater financial disclosure, stricter internal corporate control, and greater corporate responsibility over financial reporting. Therefore there is a well-established precedent of government intervention as a means of ensuring greater transparency in industries that are deemed to be insufficiently transparent.

FDA adverse event reporting. While transparency often aims to provide baseline understanding and information, sometimes further transparency is necessary in light of unexpected circumstances. In the context of drugs, the FDA implements an adverse event reporting system (FAERS) as a "database that contains information on adverse event and medication error reports submitted to FDA. The database is designed to support the FDA's post-marketing safety surveillance program for drug and therapeutic biologic products" (Food and Administration 2023, 2021). As of September 2023, there are more than 27 million reports, with the FDA receiving more than one million reports annually since 1969 (Food and Administration 2023); the FAERS data is made available to wide range of stakeholders (e.g. consumers, healthcare professionals, researchers).

Reports are voluntarily submitted by healthcare providers (e.g. physicians, pharmacists, nurses) and consumers (e.g. patients, family members, lawyers); by law, product manufacturers must relay these reports to the FDA (Food and Administration 2018). Reports are circulated and may trigger subsequent actions (e.g. evaluation by clinical reviewers in the Center for Drug Evaluation and Research) to survey post-market drug safety. Overall, the open availability of FAERS data improves awareness of drug adverse events, though it may be prone to improper interpretation without appropriate consideration for statistical validity (Kumar 2018). In comparison to the recurring, comprehensive, and proactive nature of social media transparency reports or financial reports, adverse event reporting systems provide more targeted transparency when interventions are (potentially) urgent. While our focus in designing transparency reports for foundation models largely emulates the former approaches, we highlight adverse event reporting as playing a potentially complementary role. In particular, we imagine that as specific harms of foundation models are documented, similar adverse event reporting systems (or the reuse of pre-existing systems) will be necessary (Guha et al. 2023; NAIAC 2023).

Discussion

Transparency functions as an instrument for advancing other objectives (e.g. greater public accountability and improved risk management). We aim to inculcate robust norms and industry standards around transparency while foundation models are still (relatively) nascent, in conjunction with government-driven disclosure requirements. Transparency is not a monolith: different aspects of transparency are more relevant for certain societal objectives and stakeholder groups than others. While in some cases the benefits of transparency arise from a single developer being more transparent, for others what is required is broader transparency from many developers to surface general trends. We step through several of our transparency indicators to articulate our theory of change regarding how increased transparency would help improve the societal impact of foundation models.

Greater transparency on data directly informs demographic biases in foundation model behavior (Abid, Farooqi, and Zou 2021; Luccioni et al. 2023; Bianchi et al. 2023) and copyright litigation surrounding model training data (e.g. NYT 2024). Transparency on labor practices enables awareness of, and collective action to address, labor conditions (Williams, Miceli, and Gebru 2022). Transparency on compute usage clarifies the costs of building frontier foundation models (Anderljung et al. 2023) and the viability of policies like licensing that restrict compute access (Kapoor and Narayanan 2023). Evaluations help concretize model capabilities (Wei et al. 2022; Bubeck et al. 2023) and risks (Bender et al. 2021; Weidinger et al. 2022), sharpening collective understanding (Bommasani et al. 2023c). And transparency on usage statistics as well as affected market sectors and geographies directly informs understanding of economic impact, innovation, and the concentration of power (Vipra and Korinek 2023; Bommasani et al. 2023b; UK CMA 2023).

While we advocate for greater transparency via transparency reports, we recognize that transparency initiatives have been subject to critique (Ananny and Crawford 2018; Bates, Kennedy, and Medina Perea 2023). Though some of these critiques regarding performative transparency on self-selected matters are mitigated by our approach (Zalnieriute 2021), other critiques about the limits of transparency to bring about substantive change persist (Hartzog 2023). We see improved transparency as a natural initial target given the demonstrated opacity of the foundation model ecosystem (Perrigo 2022; Hao and Seetharaman 2023; Bommasani et al. 2023a); other changes will need to follow to achieve better societal outcomes. Further, social media transparency reporting demonstrates that transparency reporting can be costly, requiring substantial investments from platforms. At present, we do not aim to factor in reporting costs, though we encourage developers to transparently discuss costs to allow policymakers and other stakeholders to better argue for cost-benefit trade-offs for transparency. For similar reasons, we also highlight the potential for Foundation Model Transparency Reports to reduce overall compliance costs for developers operating across multiple jurisdictions by reducing duplicative effort.

Broad consensus exists for improved transparency. The history of social media illustrates both the harms of pervasive opacity and the potential for institutionalized transparency. We envisage Foundation Model Transparency Reports as the structured interface for communicating information from foundation model developers to the public to benefit diverse stakeholders.

Acknowledgements

We thank Dan Ho, Daphne Keller, and Nate Persily for feedback and discussions that informed this work. This work was supported in part by the AI2050 program at Schmidt Futures (Grant G-22-63429).

References

2010. Greater transparency around government requests. https://googleblog.blogspot.com/2010/04/greater-transparency-around-government.html.

2014. 2014 Transparency Report. https://extfiles.etsy.com/Press/reports/Etsy_TransparencyReport_2014.pdf.

Abdo, A.; Krishnan, R.; Krent, S.; Welber Falcón, E.; and Woods, A. K. 2022. A Safe Harbor for Platform Research. https://knightcolumbia.org/content/a-safe-harbor-for-platform-research.

Abid, A.; Farooqi, M.; and Zou, J. 2021. Persistent anti-muslim bias in large language models. *arXiv preprint arXiv:2101.05783*.

Access Now. 2023. Transparency Reporting Index. https://www.accessnow.org/campaign/transparency-reporting-index/.

American Civil Liberties Union. 2010. Internal Report Finds Flagrant National Security Letter Abuse By FBI. https://www.aclu.org/press-releases/internal-report-finds-flagrant-national-security-letter-abuse-fbi.

Ananny, M.; and Crawford, K. 2018. Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability. *New Media & Society*, 20(3): 973–989.

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. 2023. Frontier AI Regulation: Managing Emerging Risks to Public Safety. arXiv:2307.03718.

Aspen Institute. 2021. Commission on Information Disorder Final Report. https://www.aspeninstitute.org/wp-content/uploads/2021/11/Aspen-Institute_Commission-on-Information-Disorder_Final-Report.pdf.

Bankston, K.; Schulman, R.; and Woolery, L. 2017. Case Study #3: Transparency Reporting. https://www.newamerica.org/in-depth/getting-internet-companies-do-right-thing/case-study-3-transparency-reporting/.

Bates, J.; Kennedy, H.; and Medina Perea, I. e. a. 2023. Socially meaningful transparency in data-based systems: reflections and proposals from practice. *Journal of Documentation*.

Bender, E. M.; and Friedman, B. 2018. Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics (TACL)*, 6: 587–604.

Bender, E. M.; Gebru, T.; McMillan-Major, A.; and Shmitchell, S. 2021. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 610–623.

Bianchi, F.; Kalluri, P.; Durmus, E.; Ladhak, F.; Cheng, M.; Nozza, D.; Hashimoto, T.; Jurafsky, D.; Zou, J.; and Caliskan, A. 2023. Easily Accessible Text-to-Image Generation Amplifies Demographic Stereotypes at Large Scale. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '23, 1493–1504. New York, NY, USA: Association for Computing Machinery. ISBN 9798400701924.

Birchall, C. 2021. *Radical secrecy: The ends of transparency in datafied America*, volume 60. U of Minnesota Press.

Bommasani, R.; Hudson, D. A.; Adeli, E.; Altman, R.; Arora, S.; von Arx, S.; Bernstein, M. S.; Bohg, J.; Bosselut, A.; Brunskill, E.; Brynjolfsson, E.; Buch, S.; Card, D.; Castellon, R.; Chatterji, N.; Chen, A.; Creel, K.; Davis, J. Q.; Demszky, D.; Donahue, C.; Doumbouya, M.; Durmus, E.; Ermon, S.; Etchemendy, J.; Ethayarajh, K.; Fei-Fei, L.; Finn, C.; Gale, T.; Gillespie, L.; Goel, K.; Goodman, N.; Grossman, S.; Guha, N.; Hashimoto, T.; Henderson, P.; Hewitt, J.; Ho, D. E.; Hong, J.; Hsu, K.; Huang, J.; Icard, T.; Jain, S.; Jurafsky, D.; Kalluri, P.; Karamcheti, S.; Keeling, G.; Khani, F.; Khattab, O.; Koh, P. W.; Krass, M.; Krishna, R.; Kuditipudi, R.; Kumar, A.; Ladhak, F.; Lee, M.; Lee, T.; Leskovec, J.; Levent, I.; Li, X. L.; Li, X.; Ma, T.; Malik, A.; Manning, C. D.; Mirchandani, S.; Mitchell, E.; Munyikwa, Z.; Nair, S.; Narayan, A.; Narayanan, D.; Newman, B.; Nie, A.; Niebles, J. C.; Nilforoshan, H.; Nyarko, J.; Ogut, G.; Orr, L.; Papadimitriou, I.; Park, J. S.; Piech, C.; Portelance, E.; Potts, C.; Raghunathan, A.; Reich, R.; Ren, H.; Rong, F.; Roohani, Y.; Ruiz, C.; Ryan, J.; Ré, C.; Sadigh, D.; Sagawa, S.; Santhanam, K.; Shih, A.; Srinivasan, K.; Tamkin, A.; Taori, R.; Thomas, A. W.; Tramèr, F.; Wang, R. E.; Wang, W.; Wu, B.; Wu, J.; Wu, Y.; Xie, S. M.; Yasunaga, M.; You, J.; Zaharia, M.; Zhang, M.; Zhang, T.; Zhang, X.; Zhang, Y.; Zheng, L.; Zhou, K.; and Liang, P. 2021. On the Opportunities and Risks of Foundation Models. *arXiv preprint arXiv:2108.07258*.

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.

Bommasani, R.; Soylu, D.; Liao, T.; Creel, K. A.; and Liang, P. 2023b. Ecosystem Graphs: The Social Footprint of Foundation Models. *ArXiv*, abs/2303.15772.

Bommasani, R.; Zhang, D.; Lee, T.; and Liang, P. 2023c. Improving Transparency in AI Language Models: A Holistic Evaluation. *Foundation Model Issue Brief Series*.

Boyd, D. 2016. Algorithmic Accountability and Transparency. Open Transcripts. Presented by danah boyd in Algorithmic Accountability and Transparency in the Digital Economy.

Brannon, V. C.; Killion, V. L.; Novak, W. K.; and Whitaker, L. P. 2023. First Amendment Limitations on Disclosure Requirements.

Bubeck, S.; Chandrasekaran, V.; Eldan, R.; Gehrke, J.; Horvitz, E.; Kamar, E.; Lee, P.; Lee, Y. T.; Li, Y.; Lundberg, S.; Nori, H.; Palangi, H.; Ribeiro, M. T.; and Zhang, Y. 2023. Sparks of Artificial General Intelligence: Early experiments with GPT-4. arXiv:2303.12712.

Budish, R.; Woolery, L.; and Bankston, K. 2016. The Transparency Reporting Toolkit: Survey & Best Practice Memos for Reporting on U.S. Government Requests for User Information. https://www.newamerica.org/oti/policy-papers/the-transparency-reporting-toolkit/.

Commission, E. 2022. The Digital Services Act: ensuring a safe and accountable online environment. *European Commission*.

Crawford, K. 2021. *The atlas of AI: Power, politics, and the planetary costs of artificial intelligence*. Yale University Press.

Crocker, A.; Gebhart, G.; Mackey, A.; Opsahl, K.; Tsukayama, H.; Williams, J. L.; and York, J. C. 2019. Who Has Your Back?

Edelson, L.; Chuang, J.; Franklin Fowler, E.; Franz, M.; and Ridout, T. N. 2021. Universal Digital Ad Transparency. In *TPRC49: The 49th Research Conference on Communication, Information and Internet Policy*. Available at SSRN: https://ssrn.com/abstract=3898214 or http://dx.doi.org/10.2139/ssrn.3898214.

European Commission. 2023. Commission launches public consultation on the Implementing Regulation on transparency reporting under the DSA.

European Council. 2024. 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.

Executive Order 14110. 2023. Executive Order on Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence.

Facebook. 2023. Facebook Transparent Reports.

Food, U.; and Administration, D. 2018. Questions and Answers on FDA's Adverse Event Reporting System (FAERS). https://www.fda.gov/drugs/surveillance/questions-and-answers-fdas-adverse-event-reporting-system-faers.

Food, U.; and Administration, D. 2021. FDA Adverse Event Reporting System (FAERS): Latest Quartely Data Files. https://catalog.data.gov/dataset/fda-adverse-event-reporting-system-faers-latest-quartely-data-files.

Food, U.; and Administration, D. 2023. FDA Adverse Event Reporting System (FAERS) Public Dashboard. https://www.fda.gov/drugs/questions-and-answers-fdas-adverse-event-reporting-system-faers/fda-adverse-event-reporting-system-faers-public-dashboard.

Fourrier, C.; Habib, N.; Launay, J.; and Wolf, T. 2023. What's going on with the Open LLM Leaderboard?

Gebru, T.; Morgenstern, J.; Vecchione, B.; Vaughan, J. W.; Wallach, H.; Iii, H. D.; and Crawford, K. 2021. Datasheets for datasets. *Communications of the ACM*, 64(12): 86–92.

Gebru, T.; Morgenstern, J.; Vecchione, B.; Vaughan, J. W.; Wallach, H.; Ill, H. D.; and Crawford, K. 2018. Datasheets for Datasets. *arXiv preprint arXiv:1803.09010*.

Ghosh, R.; and Faxon, H. O. 2023. Smart corruption: Satirical strategies for gaming accountability. *Big Data & Society*, 10(1): 20539517231164119.

Gorwa, R.; and Ash, T. G. 2020. *Democratic Transparency in the Platform Society*, 286–312. SSRC Anxieties of Democracy. Cambridge University Press.

Gray, M. L.; and Suri, S. 2019. *Ghost work: How to stop Silicon Valley from building a new global underclass*. Eamon Dolan Books.

Greenwald, G. 2013. NSA collecting phone records of millions of Verizon customers daily. https://www.theguardian.com/world/2013/jun/06/nsa-phone-records-verizon-court-order.

Group of Seven. 2023. Hiroshima Process International Code of Conduct for Organizations Developing Advanced AI Syste.

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. 2023. AI Regulation Has Its Own Alignment Problem: The Technical and Institutional Feasibility of Disclosure, Registration, Licensing, and Auditing. *George Washington Law Review, Symposium on Legally Disruptive Emerging Technologies*.

Han, B.-C. 2015. *The transparency society*. Stanford University Press.

Hao, K.; and Seetharaman, D. 2023. Cleaning Up ChatGPT Takes Heavy Toll on Human Workers. *The Wall Street Journal*. Photographs by Natalia Jidovanu.

Hartzog, W. 2023. Oversight of A.I.: Legislating on Artificial Intelligence. Prepared Testimony and Statement for the Record before the U.S. Senate Committee on the Judiciary, Subcommittee on Privacy, Technology, and the Law.

Hendrycks, D.; Burns, C.; Basart, S.; Zou, A.; Mazeika, M.; Song, D.; and Steinhardt, J. 2021. Measuring massive multitask language understanding. In *International Conference on Learning Representations (ICLR)*.

ISED Canada. 2023. Voluntary Code of Conduct on the Responsible Development and Management of Advanced Generative AI Systems.

Kapoor, S.; and Narayanan, A. 2023. *Licensing is neither feasible nor effective for addressing AI risks*.

Keller, D. 2021. Some Humility About Transparency. https://cyberlaw.stanford.edu/blog/2021/03/some-humility-about-transparency.

Keller, D. 2022. Hearing on Platform Transparency: Understanding the Impact of Social Media. Technical report, United States Senate Committee on the Judiciary, Subcommittee on Privacy, Technology and the Law. Statement of Daphne Keller, Stanford University Cyber Policy Center.

Kessel, J. 2016. Advancing #transparency with more insightful data. https://blog.twitter.com/official/en_us/a/2016/advancing-transparency-with-more-insightful-data.html.

Kumar, A. 2018. The Newly Available FAERS Public Dashboard: Implications for Health Care Professionals.

Lazar, S. 2023. Governing the Algorithmic City. *Tanner Lectures*.

Le Scao, T.; Fan, A.; Akiki, C.; Pavlick, E.; Ilić, S.; Hesslow, D.; Castagné, R.; Luccioni, A. S.; Yvon, F.; Gallé, M.; Tow, J.; Rush, A. M.; Biderman, S.; Webson, A.; Ammanamanchi, P. S.; Wang, T.; Sagot, B.; Muennighoff, N.; del Moral, A. V.; Ruwase, O.; Bawden, R.; Bekman, S.; McMillan-Major, A.; Beltagy, I.; Nguyen, H.; Saulnier, L.; Tan, S.; Suarez, P. O.; Sanh, V.; Laurençon, H.; Jernite, Y.; Launay, J.; Mitchell, M.; Raffel, C.; Gokaslan, A.; Simhi, A.; Soroa, A.; Aji, A. F.; Alfassy, A.; Rogers, A.; Nitzav, A. K.; Xu, C.; Mou, C.; Emezue, C.; Klamm, C.; Leong, C.; van Strien, D.; Adelani, D. I.; Radev, D.; Ponferrada, E. G.; Levkovizh, E.; Kim, E.; Natan, E. B.; De Toni, F.; Dupont, G.; Kruszewski, G.; Pistilli, G.; Elsahar, H.; Benyamina, H.; Tran, H.; Yu, I.; Abdulmumin, I.; Johnson, I.; Gonzalez-Dios, I.; de la Rosa, J.; Chim, J.; Dodge, J.; Zhu, J.; Chang, J.; Frohberg, J.; Tobing, J.; Bhattacharjee, J.; Almubarak, K.; Chen, K.; Lo, K.; Von Werra, L.; Weber, L.; Phan, L.; allal, L. B.; Tanguy, L.; Dey, M.; Muñoz, M. R.; Masoud, M.; Grandury, M.; Šaško, M.; Huang, M.; Coavoux, M.; Singh, M.; Jiang, M. T.-J.; Vu, M. C.; Jauhar, M. A.; Ghaleb, M.; Subramani, N.; Kassner, N.; Khamis, N.; Nguyen, O.; Espejel, O.; de Gibert, O.; Villegas, P.; Henderson, P.; Colombo, P.; Amuok, P.; Lhoest, Q.; Harliman, R.; Bommasani, R.; López, R. L.; Ribeiro, R.; Osei, S.; Pyysalo, S.; Nagel, S.; Bose, S.; Muhammad, S. H.; Sharma, S.; Longpre, S.; Nikpoor, S.; Silberberg, S.; Pai, S.; Zink, S.; Torrent, T. T.; Schick, T.; Thrush, T.; Danchev, V.; Nikoulina, V.; Laippala, V.; Lepercq, V.; Prabhu, V.; Alyafeai, Z.; Talat, Z.; Raja, A.; Heinzerling, B.; Si, C.; Salesky, E.; Mielke, S. J.; Lee, W. Y.; Sharma, A.; Santilli, A.; Chaffin, A.; Stiegler, A.; Datta, D.; Szczechla, E.; Chhablani, G.; Wang, H.; Pandey, H.; Strobelt, H.; Fries, J. A.; Rozen, J.; Gao, L.; Sutawika, L.; Bari, M. S.; Al-shaibani, M. S.; Manica, M.; Nayak, N.; Teehan, R.; Albanie, S.; Shen, S.; Ben-David, S.; Bach, S. H.; Kim, T.; Bers, T.; Fevry, T.; Neeraj, T.; Thakker, U.; Raunak, V.; Tang, X.; Yong, Z.-X.; Sun, Z.; Brody, S.; Uri, Y.; Tojarieh, H.; Roberts, A.; Chung, H. W.; Tae, J.; Phang, J.; Press, O.; Li, C.; Narayanan, D.; Bourfoune, H.; Casper, J.; Rasley, J.; Ryabinin, M.; Mishra, M.; Zhang, M.; Shoeybi, M.; Peyrounette, M.; Patry, N.; Tazi, N.; Sanseviero, O.; von Platen, P.; Cornette, P.; Lavallée, P. F.; Lacroix, R.; Rajbhandari, S.; Gandhi, S.; Smith, S.; Requena, S.; Patil, S.; Dettmers, T.; Baruwa, A.; Singh, A.; Cheveleva, A.; Ligozat, A.-L.; Subramonian, A.; Névéol, A.; Lovering, C.; Garrette, D.; Tunuguntla, D.; Reiter, E.; Taktasheva, E.; Voloshina, E.; Bogdanov, E.; Winata, G. I.; Schoelkopf, H.; Kalo, J.-C.; Novikova, J.; Forde, J. Z.; Clive, J.; Kasai, J.; Kawamura, K.; Hazan, L.; Carpuat, M.; Clinciu, M.; Kim, N.; Cheng, N.; Serikov, O.; Antverg, O.; van der Wal, O.; Zhang, R.; Zhang, R.; Gehrmann, S.; Pais, S.; Shavrina, T.; Scialom, T.; Yun, T.; Limisiewicz, T.; Rieser, V.; Protasov, V.; Mikhailov, V.; Pruksachatkun, Y.; Belinkov, Y.; Bamberger, Z.; Kasner, Z.; Rueda, A.; Pestana, A.; Feizpour, A.; Khan, A.; Faranak, A.; Santos, A.; Hevia, A.; Unldreaj, A.; Aghagol, A.; Abdollahi, A.; Tammour, A.; HajiHosseini, A.; Behroozi, B.; Ajibade, B.; Saxena, B.; Ferrandis, C. M.; Contractor, D.; Lansky, D.; David, D.; Kiela, D.; Nguyen, D. A.; Tan, E.; Baylor, E.; Ozoani, E.; Mirza, F.; Ononiwu, F.; Rezanejad, H.; Jones, H.; Bhattacharya, I.; Solaiman, I.; Sedenko, I.; Nejadgholi, I.; Passmore, J.; Seltzer, J.; Sanz, J. B.; Fort, K.; Dutra, L.; Samagaio, M.; Elbadri, M.; Mieskes, M.; Gerchick, M.; Akinlolu, M.; McKenna, M.; Qiu, M.; Ghauri, M.; Burynok, M.; Abrar, N.; Rajani, N.; Elkott, N.; Fahmy, N.; Samuel, O.; An, R.; Kromann, R.; Hao, R.; Alizadeh, S.; Shubber, S.; Wang, S.; Roy, S.; Viguier, S.; Le, T.; Oyebade, T.; Le, T.; Yang, Y.; Nguyen, Z.; Kashyap, A. R.; Palasciano, A.; Callahan, A.; Shukla, A.; Miranda-Escalada, A.; Singh, A.; Beilharz, B.; Wang, B.; Brito, C.; Zhou, C.; Jain, C.; Xu, C.; Fourrier, C.; Periñán, D. L.; Molano, D.; Yu, D.; Manjavacas, E.; Barth, F.; Fuhrimann, F.; Altay, G.; Bayrak, G.; Burns, G.; Vrabec, H. U.; Bello, I.; Dash, I.; Kang, J.; Giorgi, J.; Golde, J.; Posada, J. D.; Sivaraman, K. R.; Bulchandani, L.; Liu, L.; Shinzato, L.; de Bykhovetz, M. H.; Takeuchi, M.; Pàmies, M.; Castillo, M. A.; Nezhurina, M.; Sänger, M.; Samwald, M.; Cullan, M.; Weinberg, M.; De Wolf, M.; Mihaljcic, M.; Liu, M.; Freidank, M.; Kang, M.; Seelam, N.; Dahlberg, N.; Broad, N. M.; Muellner, N.; Fung, P.; Haller, P.; Chandrasekhar, R.; Eisenberg, R.; Martin, R.; Canalli, R.; Su, R.; Su, R.; Cahyawijaya, S.; Garda, S.; Deshmukh, S. S.; Mishra, S.; Kiblawi, S.; Ott, S.; Sang-aroonsiri, S.; Kumar, S.; Schweter, S.; Bharati, S.; Laud, T.; Gigant, T.; Kainuma, T.; Kusa, W.; Labrak, Y.; Bajaj, Y. S.; Venkatraman, Y.; Xu, Y.; Xu, Y.; Xu, Y.; Tan, Z.; Xie, Z.; Ye, Z.; Bras, M.; Belkada, Y.; and Wolf, T. 2022. BLOOM: A 176B-Parameter Open-Access Multilingual Language Model.

Llansó, E.; and Vogus, C. 2021. Transparency Reports. https://cdt.org/wp-content/uploads/2022/01/2021-12-20-FX-Transparency-Framework-brief-Transparency-Reports-final.pdf.

Luccioni, A. S.; Akiki, C.; Mitchell, M.; and Jernite, Y. 2023. Stable Bias: Analyzing Societal Representations in Diffusion Models. arXiv:2303.11408.

Luccioni, A. S.; Viguier, S.; and Ligozat, A.-L. 2022. Estimating the Carbon Footprint of BLOOM, a 176B Parameter Language Model. *ArXiv*, abs/2211.02001.

MacKinnon, R.; Brouillette, A.; Gutermuth, L.; Reed, L.; Maréchal, N.; Wessenauer, V.; Abrougui, A.; Cabral, S.; Sperling, I.; Rogoff, Z.; and Moore, E. 2019. 2019 RDR Corporate Accountability Index.

Meinhardt, C.; Lawrence, C. M.; Gailmard, L. A.; Zhang, D.; Bommasani, R.; Kosoglu, R.; Henderson, P.; Wald, R.; and Ho, D. E. 2023. By the Numbers: Tracking The AI Executive Order.

Miller, G. 2023. Tracking the First Digital Services Act Transparency Reports. https://www.techpolicy.press/tracking-the-first-digital-services-act-transparency-reports/.

Mitchell, M.; Wu, S.; Zaldivar, A.; Barnes, P.; Vasserman, L.; Hutchinson, B.; Spitzer, E.; Raji, I. D.; and Gebru, T. 2018. Model Cards for Model Reporting. *Proceedings of the Conference on Fairness, Accountability, and Transparency*.

Mittelstadt, B. 2019. Principles alone cannot guarantee ethical AI. *Nature Machine Intelligence*, 1(11): 501–507.

NAIAC. 2023. RECOMMENDATION: Improve Monitoring of Emerging Risks from AI through Adverse Event Reporting.

Narayanan, A.; and Kapoor, S. 2023. Generative AI companies must publish transparency reports.

NYT. 2024. THE NEW YORK TIMES COMPANY v. MICROSOFT CORPORATION, OPENAI, INC., OPENAI LP, OPENAI GP, LLC, OPENAI, LLC, OPENAI OPCO LLC, OPENAI GLOBAL LLC, OAI CORPORATION, LLC, and OPENAI HOLDINGS, LLC.

OpenAI. 2023. GPT-4 Technical Report. arXiv:2303.08774.

Patterson, D.; Gonzalez, J.; Le, Q.; Liang, C.; Munguia, L.-M.; Rothchild, D.; So, D.; Texier, M.; and Dean, J. 2021. Carbon emissions and large neural network training. *arXiv preprint arXiv:2104.10350*.

Perino, M. 2010. *The Hellhound of Wall Street: How Ferdinand Pecora's Investigation of the Great Crash Forever Changed American Finance*. Penguin Publishing Group. ISBN 9780143120032.

Perrigo, B. 2022. Exclusive: OpenAI Used Kenyan Workers on Less Than 2 Per Hour to Make ChatGPT Less Toxic. *Time*.

Pichai, S.; and Hassabis, D. 2023. Introducing Gemini: our largest and most capable AI model.

Raji, I. D.; and Buolamwini, J. 2019. Actionable Auditing: Investigating the Impact of Publicly Naming Biased Performance Results of Commercial AI Products. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '19, 429–435. New York, NY, USA: Association for Computing Machinery. ISBN 9781450363242.

Rydzak, J. 2023. The Stalled Machines of Transparency Reporting. https://carnegieendowment.org/2023/11/29/stalled-machines-of-transparency-reporting-pub-91085.

Santa Clara Principles. 2023. The Santa Clara Principles: On Transparency and Accountability in Content Moderation. https://santaclaraprinciples.org/.

Schatz, A. 2006. Tech Firms Defend China Web Policies. https://www.wsj.com/articles/SB114002162437674809.

Schneider, J.-P.; Siegrist, K.; and Oles, S. 2023. Collaborative Governance of the EU Digital Single Market established by the Digital Services Act. *University of Luxembourg Law Research Paper*, 2023(09).

Secutiries, U.; and Commision, E. 2024a. About the SEC. https://www.sec.gov/strategic-plan/about.

Secutiries, U.; and Commision, E. 2024b. Form 10-K. https://www.investor.gov/introduction-investing/investing-basics/glossary/form-10-k.

Secutiries, U.; and Commision, E. 2024c. Form 8-K. https://www.investor.gov/introduction-investing/investing-basics/glossary/form-8-k.

Secutiries, U.; and Commision, E. 2024d. Generally Accepted Accounting Principles (GAAP). https://www.investor.gov/introduction-investing/investing-basics/glossary/generally-accepted-accounting-principles-gaap.

Secutiries, U.; and Commision, E. 2024e. Generally Accepted Accounting Principles (GAAP). https://www.investor.gov/introduction-investing/investing-basics/glossary/generally-accepted-accounting-principles-gaap.

Stoughton, K.; and Rosenzweig, P. 2022. Toward Greater Content Moderation Transparency Reporting. Lawfare.

Trust and Safety Professional Association. 2023. Transparency Reporting. https://www.tspa.org/curriculum/ts-fundamentals/transparency-report/.

UK CMA. 2023. AI Foundation Models: Initial Report.

United States Congress. 2023. AI Foundation Model Transparency Act.

Urman, A.; and Makhortykh, M. 2023. How transparent are transparency reports? Comparative analysis of transparency reporting across online platforms. *Telecommunications Policy*, 47(3): 102477.

Vermeulen, M. 2021. The Keys to the Kingdom.

Vipra, J.; and Korinek, A. 2023. Market concentration implications of foundation models: The Invisible Hand of ChatGPT. *The Brookings Institution*.

Vogus, C.; and Llansó, E. 2021. Making Transparency Meaningful: A Framework for Policymakers. *Center for Democracy and Technology*.

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. *Transactions on Machine Learning Research*. Survey Certification.

Weidinger, L.; Uesato, J.; Rauh, M.; Griffin, C.; Huang, P.-S.; Mellor, J.; Glaese, A.; Cheng, M.; Balle, B.; Kasirzadeh, A.; Biles, C.; Brown, S.; Kenton, Z.; Hawkins, W.; Stepleton, T.; Birhane, A.; Hendricks, L. A.; Rimell, L.; Isaac, W.; Haas, J.; Legassick, S.; Irving, G.; and Gabriel, I. 2022. Taxonomy of Risks Posed by Language Models. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, 214–229. New York, NY, USA: Association for Computing Machinery. ISBN 9781450393522.

White House. 2023. Ensuring Safe, Secure, and Trustworthy AI.

Williams, A.; Miceli, M.; and Gebru, T. 2022. *De Anima: On the Soul*.

World Federation of Advertisers. 2020. WFA and platforms make major progress to address harmful content. https://wfanet.org/knowledge/item/2020/09/23/WFA-and-platforms-make-major-progress-to-address-harmful-content.

X. 2023. An update on Twitter Transparency Reporting. https://blog.twitter.com/en_us/topics/company/2023/an-update-on-twitter-transparency-reporting.

Zalnieriute, M. 2021. "Transparency-Washing" in the Digital Age : A Corporate Agenda of Procedural Fetishism. Technical report.

Zhu, J. 2015. A perfect EFF score! We're proud to have your back. https://wordpress.com/blog/2015/06/17/a-perfect-eff-score-were-proud-to-have-your-back/.