Tracing the Evolution of Information Transparency for OpenAI's GPT Models through a Biographical Approach

Zhihan Xu, Eni Mustafaraj

Department of Computer Science
Wellesley College
zx101@wellesley.edu, emustafa@wellesley.edu

Abstract

Information transparency, the open disclosure of information about models, is crucial for proactively evaluating the potential societal harm of large language models (LLMs) and developing effective risk mitigation measures. Adapting the biographies of artifacts and practices (BOAP) method (Hyysalo, Pollock, and Williams 2019) from science and technology studies, this study analyzes the evolution of information transparency within OpenAI's Generative Pre-trained Transformers (GPT) model reports and usage policies from its inception in 2018 to GPT-4, one of today's most capable LLMs. To assess the breadth and depth of transparency practices, we develop a 9-dimensional, 3-level analytical framework to evaluate the comprehensiveness and accessibility of information disclosed to various stakeholders. Findings suggest that while model limitations and downstream usages are increasingly clarified, model development processes have become more opaque. Transparency remains minimal in certain aspects, such as model explainability and real-world evidence of LLM impacts, and the discussions on safety measures such as technical interventions and regulation pipelines lack in-depth details. The findings emphasize the need for enhanced transparency to foster accountability and ensure responsible technological innovations.

Introduction

Abstraction is a fundamental concept in computing that simplifies complex systems by removing their internal details and the origins of their inputs and outputs (Selbst et al. 2019). Just like pre-made foods offer convenience and simplicity by abstracting away the raw ingredients and the cooking procedures, large language models (LLMs) transform the software development ecosystem by enabling more efficient development processes, where AI practitioners can leverage the advanced capabilities of existing models, such as GPT-4, to build wide-ranging applications, without access to model details such as their intricate inner workings and sources of training data.

However, this abstraction also breeds opacity and can thus lead to serious implications. As more state-of-the-art downstream models and applications adapt from only a few large language models, these models become simultaneously the single source of truth and the single point of failure. For instance, the inherent biases within these models could be passed on to all these downstream systems, potentially amplifying societal harms across domains (Bommasani et al. 2021). Therefore, information transparency, the open sharing of information about these models, is essential for proactively evaluating and mitigating relevant risks.

Despite its importance, transparency for LLMs has not yet been widely discussed in academic publications and public discourse (Liao and Vaughan 2023; Masotina, Musi, and Spagnolli 2023). This study adapts the biographies of artifacts and practices (BOAP) method (Hyysalo, Pollock, and Williams 2019) from science and technology studies to evaluate how transparency for OpenAI's GPT models, one of the world's leading large language models, has changed over time. Through a biographical lens, this research aims to situate the changes in transparency practices within the broader technological, organizational, and socio-operational contexts to understand how various factors have influenced transparency for emerging technologies like GPTs.

The central research question this paper aims to answer is, how has the information transparency regarding OpenAI's GPT models evolved? This question can be further broken down into two smaller research questions (RQs) as follows:

- RQ1 (Breadth of Transparency): What categories of information (e.g., governance structures, model limitations, applications) has OpenAI disclosed regarding their GPT models, and how has the scope of this disclosed information evolved?
- RQ2 (Depth of Transparency): How has OpenAI changed their approach to communicating information within each category over time, in terms of the comprehensiveness, accessibility, and underlying social awareness demonstrated by the disclosed information?

This work has three main contributions. First, it proposes a concise transparency framework to qualitatively assess the breadth and depth of information disclosure for large language models and other emerging technologies. Second, it introduces a biographical perspective to the study of transparency changes over time, offering a more comprehensive interpretation within the sociotechnical context compared to existing approaches. Lastly, this paper translates technical model reports into an accessible historical narrative to facil-

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

itate interdisciplinary collaboration, helping a wider range of stakeholders from technical and non-technical communities engage in the governance of large language models.

Background and Related Works

Overview of LLMs and GPTs

Large language models are deep neural networks trained on vast amounts of diverse data to understand and generate natural language text. These models have grown exponentially in size in recent years, with the definition of "largeness" evolving as the models continue to scale. While GPT-2, with its 1.5 billion parameters, was first referred to as a large-scale language model (Solaiman et al. 2019), today's LLMs have reached hundreds of billions and even trillions of parameters (Du et al. 2022).

Typically, LLMs are pre-trained with self-supervised learning techniques to obtain general language skills, and then fine-tuned on smaller, task-specific datasets to tailor their capabilities for specialized downstream applications, such as text classification and machine translation. These fine-tuned models can be further deployed into LLM-infused applications across various domains to be consumed by end-users. ChatGPT, for instance, is a popular LLM-infused chatbot built upon the GPT series of models.

The pre-trained, general-purpose nature of prominent LLMs classifies them as "foundation models" – versatile models that can be adapted efficiently to wide-ranging downstream use cases with minimal additional training on specialized datasets (Bommasani et al. 2021). The concept is analogous to crude oil as a fundamental raw material. Like crude oil is refined into various petroleum products including gasoline, lubricants, and plastics to power different technologies and industries, foundation large language models provide core language capabilities that can be tailored to create diverse AI applications with less data and compute.

The LLM ecosystem involves several stakeholder groups: the foundation model providers who develop and distribute the pre-trained LLMs, the application developers who adapt these models for downstream use, the end users who consume the LLM-infused applications, and the impacted groups directly or indirectly affected by these applications' outputs (Bommasani et al. 2023; Hacker, Engel, and Mauer 2023; Liao and Vaughan 2023).

OpenAI and GPTs GPT models are one of the most powerful and well-known large language models, developed by OpenAI, an artificial intelligence research laboratory. OpenAI was initially founded as a non-profit organization in 2015 and later transitioned into a "capped-profit" company in 2019 to secure more research funding. In 2018, OpenAI introduced the first version of GPT. Trained on a diverse corpus of unlabeled text data using unsupervised learning, the model demonstrated the potential of pre-training for transfer learning to various natural language processing (NLP) tasks (Radford et al. 2018). Subsequently, they released GPT-2 in 2019, GPT-3 in 2020, and GPT-4 in 2023, with each model version exhibiting more advanced language capabilities.

BOAP and Technology Biographies

A biographical approach to technology studies refers to understanding the development of technologies within their broader social and cultural contexts, examining their life cycle from historical and evolutionary perspectives. As information technologies become more integrated into social practices, scholars use software biographies to document the technological changes in algorithms, architecture, and operations and interpret how they are shaped by social factors like organizational dynamics and stakeholder requirements (Glaser, Pollock, and D'Adderio 2021; Helmond, Nieborg, and van der Vlist 2019).

BOAP is a framework that explores the complex interplay between technology and society, tracing different phases of sociotechnical innovations from inception to impact. Its key principles advocate for a comprehensive spatiotemporal scope that transcends traditional "snapshot" studies and an examination of the broader ecologies of interconnected actors surrounding the technologies and practices (Hyysalo, Pollock, and Williams 2019). Applying the BOAP approach, for instance, Wiegel (2016) conducted a longitudinal case study to analyze how strategic planning software was developed and used over time as a sociotechnical innovation in the automotive industry.

Facing the opacity of large language models, many recognize the value of evolutionary perspectives on auditing model updates (Chen, Zaharia, and Zou 2023) or practice changes (Bommasani et al. 2023) in driving governance and regulation progress. Biographies, however, have a unique advantage in accounting for broader contexts and implications of evolutions in technological artifacts and practices, thus inviting deeper reflections on innovation processes. This paper adapts the BOAP framework to analyze GPT, one of the most representative LLMs, from its inception in 2018 to today's massive-scale GPT-4. It aims to explore transparency changes within the sociotechnical contexts surrounding the evolution of GPT models.

Transparency

Transparency denotes disclosing information to support consumers in making decisions (Turilli and Floridi 2009). It is widely related to concepts of accountability, openness, honesty, and integrity (Ball 2009; Higgins and Tang 2024). Lack of transparency in organizational practices may conceal malpractice, ultimately resulting in corporate scandals and extensive societal harm (Bommasani et al. 2023).

Transparency is fundamental to the safety, reliability, and fairness of emerging technologies like AI (Bommasani et al. 2023; Mökander et al. 2023; Liao and Vaughan 2023). Over the years, scholars and regulatory bodies have called for improving transparency in the field. Researchers have proposed information communication mechanisms from model cards that report intended use cases and relevant performance metrics (Mitchell et al. 2019) to datasheets that document the curation and consumption of datasets (Gebru et al. 2021). Regarding regulatory initiatives, transparency emerges as the most commonly referenced principle across global AI ethics guidelines (Jobin, Ienca, and Vayena 2019).

As a trending AI technology, large language models have faced criticism for a lack of transparency. Bommasani et al. (2023) developed the Foundation Model Transparency Index as a comprehensive rubric to quantify transparency of the most influential model providers, aiming to push for progress on transparency practices over time. Liao and Vaughan (2023) summarized the sociotechnical factors obstructing transparency for LLMs, including stakeholder and application complexity, proprietary nature and competitive landscape, technical incomprehensibility due to large-scale architecture, and unpredictable model behaviors. Drawing from prior research in AI transparency, the authors suggested a goal-oriented approach for developing and evaluating transparency practices, meaning considering stakeholders' diverse information needs based on their objectives.

Limitations of Transparency Transparency in foundation large language models has several limitations. To start with, it is unrealistic to expect full transparency from the model providers, as they may be unwilling to disclose all model details to protect intellectual property rights and maintain competitive advantages (Bommasani et al. 2023). Besides, transparency alone does not guarantee ethical outcomes, as releasing misleading or inappropriate information can impair accountability and safety. For example, partial details about personal data handling can create a misperception of security and privacy that hinders accountability (Turilli and Floridi 2009). Moreover, excessive openness could be exploited by malicious actors to identify vulnerabilities or manipulate the model, leading to societal harm and unintended consequences.

Data and Methods

Guided by the BOAP framework, we included various data types across time from 2018 to 2023 to enrich the spatiotemporal diversity of GPT model information. Additionally, we contextualized the evolution of GPT models and related transparency practices within interconnected ecologies, ranging from stakeholder groups across the LLM ecosystem to broader information consumers like policymakers, regulators, and AI researchers. We specified nine transparency domains and three depth levels to encompass multiple facets of the LLM ecosystem and broader perspectives of information consumers.

Then, we open coded the selected artifacts against this pre-defined transparency framework: we labeled the content of each artifact based on the coverage of specific transparency domain(s), and evaluated the depth level at which information was disclosed within each covered domain. This open coding allows us to identify patterns in how OpenAI expanded or shrank the scope and granularity of their transparency practices over time across different artifacts.

Selected Artifacts

The raw data comprises model documentation and usage policies for OpenAI's GPT series of models.

Model Documentations This study includes 9 artifacts spanning 4 model versions (Table 1). It incorporates technical reports introducing each model – GPT-1 (Radford et al. 2018), GPT-2 (Radford et al. 2019), GPT-3 (Brown et al. 2020), and GPT-4 (OpenAI 2023) – and available model cards for GPT-2 (OpenAI 2019) and GPT-3 (OpenAI 2020) on GitHub. The purpose of technical reports and model cards is to communicate critical information about the models, such as their development processes, performance, limitations, and use cases.

Model	Title	Version
GPT-1	Improving Language Understanding by Generative Pre-Training	2018
GPT-2	Language Models are Unsupervised Multitask Learners	2019
GPT-2	Model Card	2019.11
GPT-2	Release Strategies and the Social Impacts of Language Models	2019.11
GPT-3	Language Models are Few-Shot Learners	2020.07
GPT-3	Model Card	2020.09
GPT-3	Understanding the Capabilities, Limitations, and Societal Impact of Large Language Models	2021.02
GPT-4	GPT-4 Technical Report	2023.12
GPT-4	GPT-4 System Card	2023.03

Table 1: A List of Selected GPT Model Documentations

For the other three supplementary documents, one paper (Solaiman et al. 2019) details GPT-2's release strategies and processes based on research conducted internally and through collaborations with external partners. Another document (Tamkin et al. 2021) features a conference discussion of GPT-3's limitations and societal impacts between OpenAI, the Stanford Institute for Human-Centered Artificial Intelligence, and other academic institutions. The GPT-4 system card (OpenAI 2023) accompanies the GPT-4 technical report, outlining the model's potential harms and implemented interventions for risk mitigation.

Usage Policies To understand OpenAI's approach to model deployment, we include multiple versions of their usage policies, which specify the models' intended use cases, monitoring mechanisms, and regulatory pipelines. These guidelines provide insights into what OpenAI views as safe and responsible use of their GPT models.

We retrieved copies of previous versions of usage policies through the Wayback Machine (Internet Archive 2001), a digital archive of web pages. As of March 11, 2024, the Wayback Machine had recorded 870 screenshots of OpenAI's usage policies between March 10, 2023 and March 11, 2024. By examining the last updated date listed on available records, we identified four distinct versions of the usage policies updated on February 15, 2023 (the earliest available version), March 17, 2023, March 23, 2023, and January 10, 2024 (the current version). For earlier versions of the usage policies unavailable at the current URL, we referenced the changelog (Figure 1) attached to the current usage policies

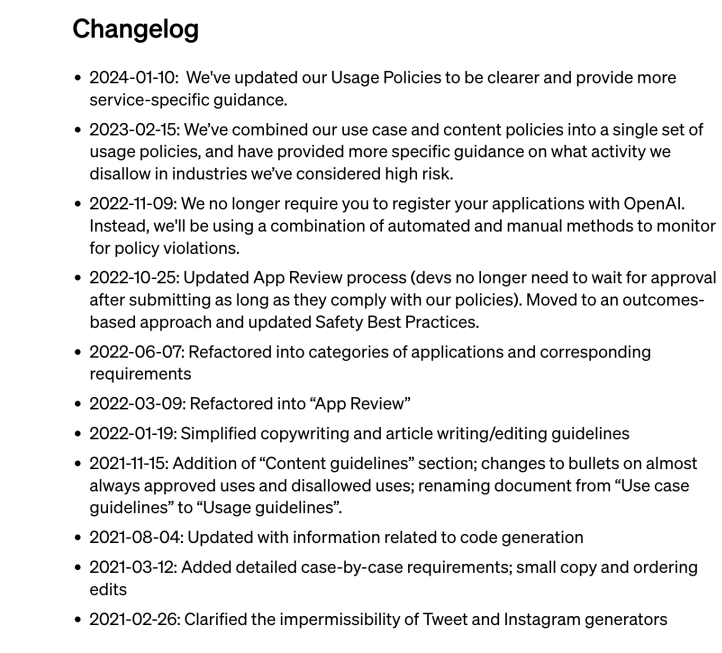

Figure 1: A Changelog Summarizing Changes Made to the OpenAI Usage Policies as of January 2024 Update

(OpenAI 2024) to identify any updates made to the document. We traced the updates across different versions of the policies over time to inspect how OpenAI's perspective on model downstream use evolved.

Transparency Framework: Breadth

To enhance LLM governance, Mökander et al. (2023) proposed auditing different entities across the LLM ecosystem, including foundation model providers, foundation models themselves, and LLM-infused downstream applications. We adopted their three dimensions of governance, model, and application as the core of our transparency framework, with slight modifications to focus on the foundation models and the impact of their transparency on the broader LLM ecosystem. This framework assesses how well the information disclosed by OpenAI serves diverse stakeholders, such as policymakers, application developers, and other interested parties. we define each aspect of transparency as follows:

- **Governance Transparency:** Disclosure of the foundation model's "making" process from development to distribution. Examining the model provider's decision-making throughout the process will support policymakers and regulators in supervising governance and improving regulatory standards.
- **Model Transparency:** Disclosure of the foundation model's limitations, explainability, and model-level risk mitigation strategies. The exposure of known model vulnerabilities and corresponding mitigation actions will facilitate model understanding and evaluations among AI researchers and practitioners.
- **Application Transparency:** Disclosure of the foundation model's intended use cases, monitoring and regulatory mechanisms, and observed or anticipated societal impact. This information will improve social understanding of LLMs and guide responsible model use among application developers and end-users.

Dimension	Subdomains
Governance Transparency	Organizational Structure, Development Process, Release Protocol
Model Transparency	Model Limitations, Model Mitigations, Model Explainability
Application Transparency	Intended Use Cases, Monitoring and Regulation, Impact Assessment

Table 2: Nine Subdomains Under the Three Main Transparency Dimensions: Governance, Model, and Application

To assess transparency from multiple facets of the LLM ecosystem, we identified nine subdomains evenly spanning across the three dimensions (Table 2). Seven of these subdomains were derived by condensing the majority of the 100 low-level transparency indicators outlined in the Foundation Model Transparency Index (Bommasani et al. 2023). Additionally, we introduced two novel subdomains: organizational structure and model explainability, to assess transparency regarding the involved social entities and the models' inner workings. This resulting framework maintains a manageable level of granularity for a close, qualitative analysis of model-related documents. Besides, it offers greater flexibility to capture valuable nuances, such as the intricate interplay between different transparency factors, thus supporting a cohesive narrative on the evolution of transparency in GPT's documentation. In the rest of the section, we will define each subdomain (Table 3) and justify their inclusions.

Organizational Structure This subdomain includes but is not limited to organizational values and incentives, roles and responsibilities of internal and external stakeholders, and governance policies that guide their decision-making processes. Selbst et al. (2019) emphasize the importance of social actors, such as incentives and decision-making cultures within an organization, in defining and promoting fairness in sociotechnical systems. Examining social actors can also reveal the exploitation of human labor in data creation, annotations, and providing feedback that the models learn from (Li et al. 2023; Arrieta-Ibarra et al. 2018; Crawford 2021; Gray and Suri 2019). Overall, information about people and practices can contribute to greater accountability (Mökander et al. 2023; Mallin 2003) and better governmental regulations (Bommasani et al. 2023).

Development Process This subdomain includes all aspects of model implementations: data sourcing, model training, and resource consumption. First, training data has been a central focus of transparency advocacy. With a new data documentation framework proposed, Gebru et al. (2021) called for transparency on the motivation, composition, collection, processing, and use of training data. About 20% of the transparency indicators selected by Bommasani et al. (2023) relate to data features listed above. Data transparency

Subdomain	Definition
Organizational Structure	Does the artifact recognize the social factors influencing the technology's development?
Development Process	Does the artifact disclose information related to the stages of model development?
Release Protocol	Does the artifact introduce the rationale and processes for model release?
Model Limitations	Does the artifact discuss the limitations and risks of the foundation model?
Model Mitigations	Does the artifact propose and evaluate any model risk mitigation strategies?
Model Explainability	Does the artifact include information about model interpretability and explainability?
Intended Use Cases	Does the artifact specify the model's intended users and application scenarios?
Monitoring & Regulation	Does the artifact explain how the model will be monitored and regulated after its release?
Impact Assessment	Does the artifact address the potential societal harms associated with model deployment?

Table 3: Definition of Nine Subdomains Used to Evaluate the Breadth of Transparency in GPT Model Reports & Usage Policies

is believed to help identify, evaluate, and mitigate risks over privacy, bias, hallucination, and copyright (Hacker, Engel, and Mauer 2023; Ferrara 2023; Weidinger et al. 2022; Bender et al. 2021; Zhang et al. 2023; Ganguli et al. 2022).

Second, while full disclosure of training details may be unrealistic due to privacy and proprietary considerations, high-level descriptions of model architecture and algorithms are encouraged (Mitchell et al. 2019). Some scholars like Liu et al. (2023) advocated for open sourcing all training components from data to code to improve transparency and reproducibility for LLMs.

Lastly, many researchers have expressed concerns over machine learning models' substantial energy and environmental costs from training to operations (Strubell, Ganesh, and McCallum 2019; Patterson et al. 2021; Lacoste et al. 2019; Schwartz et al. 2020; Luccioni and Hernandez-Garcia 2023). Transparency in resource consumption can inform policy decisions to reduce sociopolitical harm and promote sustainable model development practices.

Release Protocol This subdomain encompasses various aspects of model release, including release strategies, distribution channels, and access levels for different stakeholders. Facing the trade-off between open release for risk investigation and restricted access for safeguards, existing model providers hold varying attitudes toward model releases. Liang et al. (2022) recommended establishing a shared standard to govern the release process of foundation models. They proposed a framework for developing release policies that covers "what to release, to whom to release, when to release, and how to release." Transparency in release decision-making and processes can facilitate the sharing of best practices and the convergence toward universally accepted release practices.

Model Limitations This subdomain indicates the model's inherent weaknesses tied to its structures, properties, and development processes. For example, large language models have intrinsically limited reliability as they are subject to hallucinations, or generating non-existent or incorrect information, which leads to potential misinformation when deployed (OpenAI 2023). Making them transparent will foster a clear understanding of the model's fundamental constraints, paving the way for the responsible use of foundation models and mitigating the downstream harm (Mökander et al. 2023).

Model Mitigations This subdomain sheds light on the risks that model providers prioritize addressing and how they define the success of their mitigation strategies. Using the regulatory requirements for social media platforms as an example, Narayanan and Kapoor (2023) emphasized the need for generative AI companies to describe and evaluate their risk mitigation mechanisms to improve the visibility of safety challenges. Transparent mitigation efforts not only demonstrate model providers' commitment to improving safety measures but also facilitate external oversight to identify potential blind spots.

Model Explainability This subdomain focuses on the evaluation and techniques for enhancing model interpretability (the understandability of the model's decisions) and explainability (the understandability of the model's inner workings) (Hrín 2023). Limited model interpretability can undermine their applications in high-stakes scenarios and impede the development of effective safety measures (Singh et al. 2024). In contrast, interpretable models can self-generate reliable explanations that accurately reflect their decision logic, thereby ensuring fairness and accountability (Rudin 2019). Effectively communicating a model's explainability enables internal and external stakeholders to more thoroughly evaluate its trustworthiness and reliability.

Intended Use Cases This subdomain specifies the model's intended users and use cases, informing stakeholders about what the model should and should not be used for to minimize inappropriate model usage (Mitchell et al. 2019). Even though the foundation model's diverse applications pose challenges to envisioning its use before deployment, Bender et al. (2021) stressed the importance of exploring potential stakeholders and use cases to understand its broader risks (Bender et al. 2021). The documentation of intended usage can inform downstream developers' design decisions and guide regulators and policymakers to enforce responsible deployment.

Monitoring and Regulation Tracking and controlling model usage are important to prevent potential misuse. A

clear statement on monitoring efficacy is considered the final piece of transparency practices for open-accessed foundation models (Bommasani et al. 2023). Transparency in monitoring and regulation mechanisms can provide insights into the efforts made by model providers to ensure safety and security and empower affected individuals and communities to seek redress in case of harm or unethical use.

Impact Assessment This subdomain assesses two categories of harms associated with deploying LLMs: observed harms, which are supported by empirical evidence, and anticipated harms, which have not yet manifested (Weidinger et al. 2022). It also examines whether the model provider has proposed or implemented any deployment-level mitigations. Previous studies have developed various frameworks for algorithmic impact assessment (Selbst 2021; Reisman et al. 2018; Mantelero 2018; Schiff et al. 2020; Solaiman et al. 2023), and emphasized the significance of identifying potential recipients – affected individuals, groups, and sectors – of harm (Bommasani et al. 2023; Mitchell et al. 2019). Transparent assessment of negative impacts will uncover model providers' potential ethical blind spots, contributing to more comprehensive risk mitigations and harm prevention.

Transparency Framework: Depth

Turilli and Floridi (2009) stated that transparency, from the view of information consumers, depends on three factors: information availability, accessibility, and usefulness. On the other hand, information providers have full control over what information to disclose and how, reflecting their ethical considerations. Based on these factors, to evaluate the depth of transparency across different domains, we delineate transparency into three levels:

- **Availability:** Does the model provider share comprehensive information in each transparency domain?
- **Accessibility:** How is this information communicated? Is it presented in a clear and understandable way?
- **Awareness:** Does the model provider demonstrate awareness of the model's societal implications and willingness to address potential harms? Do they share useful information for stakeholders' decision-making process?

The following example illustrates how we evaluate transparency depth regarding model limitations using this framework: Availability entails providing a comprehensive description and evaluation of model limitations. Accessibility involves clearly explaining limitations using concrete examples, infographics, and a well-organized structure to facilitate understanding for information consumers. Awareness can be conveyed through discussing the relevant societal implications of these limitations, such as potential risks or harms to different stakeholder groups. It also includes presenting thorough justification (e.g., citing legitimate safety considerations) if any limitations information is withheld.

Findings

This section presents our findings from analyzing GPT model documents using the transparency framework defined earlier. To answer RQ1, we summarize how the scope of disclosed information has evolved over time for each of the nine transparency domains. For RQ2, we analyze how the depth of information has changed within each dimension. The findings are organized chronologically by model version to form a biography of the GPT models.

Governance Transparency

Over time, GPT model reports exhibited contrasting trends in transparency across three domains under governance transparency. While transparency increased for organizational structure through detailed roles and responsibilities, there was a notable decrease in transparency regarding the model development process and release protocol.

Organizational Structure The GPT model reports demonstrate increasing transparency regarding the roles and responsibilities of people involved in model development and deployment. While GPT-1 only listed authors of the research paper, GPT-2 acknowledged the participation of more parties such as external collaborators and data laborers. GPT-3 included a 1-page formal contributions section, identifying key contributors with high-level details about their duties, such as implementing training infrastructure and conducting training data analysis. With a 3-page expanded credits section, GPT-4 further specified the distribution of work within the organization by the stages of the development pipeline, including pre-training, evaluation and analysis, and deployment. They also clarified the role of specific individuals, such as the data collection lead, involved in each stage.

This structured and comprehensive contributions section made the organizational structures more accessible, providing greater visibility into how the work has been divided and who can potentially be held accountable. This evolution indicates improved transparency regarding the social factors in GPT development and uncovers an expansion of project scope and an increasingly complex and fragmented model development pipeline.

However, transparency gaps in social factors still exist. For example, the GPT-4 report does not provide details on the use of data labor, despite its disclosure of reliance on the Reinforcement Learning from Human Feedback (RLHF) technique that requires human annotation of sensitive content. Vaguely referring to data annotators as "vendor-managed workers" without explicitly naming specific vendor partners, OpenAI did not address existing concerns over labor exploitation, such as their contract with Sama for outsourcing data labeling of toxic content (Perrigo 2023). Additionally, the reports disclosed little information related to OpenAI's governance framework, decision-making processes, and organizational incentives.

Development Process Improved transparency was observed in data sourcing, training procedures, and model evaluations from GPT-1 to GPT-3. Specifically, OpenAI provided more thorough descriptions and illustrative examples to enhance the availability and accessibility of information about the model development process.

For GPT-1, information was disclosed mostly through citing public datasets and benchmarks. For GPT-2, OpenAI introduced their self-curated web dataset by detailing their data sourcing process and giving an overview of the data compositions. However, they did not make the full training dataset publicly available. Both GPT-1 and GPT-2 provided sufficient details about model architectures and training approaches, effectively contributing to the transparency and reproducibility of their research.

For GPT-3, OpenAI presented extensive details to explain their increasingly complex model evaluation procedures. For instance, they provided nuanced samples showing the phrasing and formatting used for prompting and fine-tuning to clarify the test settings. They also released novel synthetic datasets they designed to evaluate GPT-3's new capabilities, like arithmetics and word scrambles, to facilitate external investigations. Moreover, with informative statistics, they revealed and compared hyperparameters, computational resource consumption, and evaluation results for model variants at different scales across different test settings.

However, OpenAI drastically reduced transparency by withholding almost all information about the training data, model size, and energy usage, for GPT-4, considering "competitive landscape and safety implications." While OpenAI planned to "make further technical details available to additional third parties," they did not identify any potential independent collaborators and only stated preliminary ideas around third-party auditing (OpenAI 2023). This huge decline in transparency from GPT-3 to GPT-4 significantly hindered reproducibility, preventing public scrutiny and external verification of the model's safety and reliability.

Release Protocol There has been a trend toward less transparency in the release of GPT models over time. GPT-1 and GPT-2 were open-sourced for free, while GPT-3 and GPT-4 were released via controlled API (Application Programming Interface) access and integrated downstream products like ChatGPT. Even the red-teaming experts who closely collaborated with OpenAI on model evaluation and risk mitigations could only access GPT-4 through controlled API, as disclosed in the GPT-4 system card.

For GPT-2, OpenAI offered a comprehensive explanation of the model release rationale, strategies, and processes. They included justifications for their staged release decision (e.g., facilitating safety research), cited evidence from interdisciplinary research and discussions, and documented the specific release timeline (Solaiman et al. 2019). Contrarily, for GPT-3 and GPT-4, OpenAI's decision-making process and criteria for model release became much more opaque. The GPT-3 release appeared motivated primarily by the model's usability rather than an explicit weighing of risks and benefits. Similarly, there was no clear specification on why and how OpenAI released GPT-4. The release of GPT-4 seemed to be taken for granted, with OpenAI focusing their communication on addressing potential issues that may arise after model deployment rather than discussing their initial decision to release the model.

Model Transparency

The GPT model reports demonstrated a general increase in model transparency, with comprehensive analysis of model limitations, descriptions of mitigations implemented to address these limitations, and recognition of the challenges and significance in explaining the model's decision logic. Nevertheless, despite these improvements, OpenAI still provided limited details about their efforts to mitigate model-level risks and improve explainability.

Model Limitations From GPT-1 to GPT-4, the discussions of model limitations became more well-structured, comprehensive, and application-oriented, suggesting the model provider's improved social awareness. Discussions around model limitations in early GPT models were brief, implicit, and scattered throughout the model evaluation section. For example, GPT-2's lack of reliability was noted as a performance-wide shortcoming in its summarization capabilities.

Starting from GPT-3, OpenAI introduced a dedicated section on model limitations about one and a half pages in length (approximately 3.6% of the model report), covering technical constraints (e.g., structural and algorithmic limitations), societal risks (e.g., inherent societal biases), and lack of model explainability (e.g., the ambiguity of machine understanding). With GPT-4, the discussion was embedded throughout a 17-page summary of observed safety challenges (approximately 36.2% of the model report), contextualizing model limitations by highlighting challenges associated with reliability (e.g., hallucination) and user-model interaction (e.g., the model being overly gullible or misleadingly confident). It also elaborated on how these limitations could potentially lead to societal harm. This comprehensive and application-oriented framing not only improved information accessibility by making technical limitations concrete and understandable, but also reflected a growing social awareness by considering their real-world implications.

Notably, the discussion around model limitations evolved along with model advancements, reflecting a deeper understanding of the complexities involved. For example, while early narratives focused on improving reliability as a straightforward objective, more recent model reports recognized the double-edged nature of this pursuit: improved reliability, though making the model more useful and truthful, can reinforce the misuse potential of language models (Brown et al. 2020) and may increase the tendency of overreliance (OpenAI 2023).

Model Mitigations GPT model reports demonstrated significant progress in addressing potential limitations and risks at the model level, shifting from no discussions of mitigations to proposed mitigations and finally to implemented mitigations. This evolution aligns with the increased transparency regarding model limitations discussed above. Identifying such limitations is the prerequisite to proactively addressing high-priority risks.

In the era of GPT-3, OpenAI first proposed potential model-level risk mitigations in a few brief bullet points, suggesting modifications to data or training processes to address

harmful model biases (Tamkin et al. 2021). However, despite emphasizing the importance of bias intervention, they offered only "brief comments on future directions" while calling for external engagement on bias mitigations (Brown et al. 2020). Overall, there was no disclosure of specific mitigations that have been planned or implemented.

With the release of GPT-4, OpenAI introduced concrete efforts to implement mitigations directly into the model. These efforts were described in a 3-page section titled "Risks & mitigations" in their technical report, with further elaboration spanning 4.5 pages of the 30-page system card. They focused on two major areas: refusal mitigations to address model-user interaction challenges and hallucination mitigations to address reliability issues. The former involved tuning the foundation model to appropriately respond to disallowed, sensitive, or normal requests, while the latter fine-tuned GPT-4 to reduce its frequency of hallucinations. The model documents described the mechanism used for these model-level mitigations at a high level and evaluated their efficacies by comparing pre and post-intervention metrics, showcasing increased transparency compared to prior model versions. However, limited details were shared about the specific implementations, for instance, the dataset used to fine-tune desired model behaviors or the criteria used to guide data annotation.

Model Explainability As GPT models exhibited diminishing explainability due to scaling, OpenAI has recognized this lack of explainability both as a limitation and a potential risk, as it may lead to unexpected model behaviors or capability jumps when deployed in real-world applications.

For early GPT models, OpenAI attempted to speculate on their internal workings from their task performance, but they did not directly address the concept of model explainability. While they initially identified the lack of interpretability and predictability of model behaviors as a limitation in communication for GPT-3, they considered this a common problem for large-scale deep learning systems (Brown et al. 2020). Their future research plan related to improving explainability focused solely on understanding why increasing the model size improves performances (Tamkin et al. 2021).

GPT-4 reports first highlighted the danger of unexpected emergent behaviors, calling for more interpretability and explainability research to open the black box (OpenAI 2023). However, we believe that relevant discussions have remained superficial, as they focus more on behavioral forecasts rather than explanations. For example, the authors substantiated the positive correlations between model performance and size, known as the scaling law (Kaplan et al. 2020), with empirical evidence to enable predictions of post-scaling model performance. Nevertheless, their findings did not delve into the underlying causes of these scaling capabilities. While predictable scaling helped forecast model performance, it did not imply predictability in all aspects of model behaviors, including the risky emergent capabilities or capability jumps within broader system dynamics.

Despite the increasing awareness of model explainability issues and their societal implications, the model reports lacked a comprehensive analysis of the model's current explainability and provided limited insights into techniques used to interpret the model's decisions and outputs. There exist opportunities for improvement in both the transparency of the model itself and the transparency regarding the efforts undertaken to bolster model explainability.

Application Transparency

The application transparency for GPT models has improved over time, as evidenced by more comprehensive and practical discussions of model usage across all three subdomains: intended use cases, monitoring and regulation mechanisms, and societal harm assessments and interventions. Additionally, OpenAI has committed to continuously updating usage policies in response to emerging risks and use cases, reflecting a growing mindset in social awareness. To enhance transparency around policy evolution, they have included a policy changelog for stakeholders to examine historical changes made to their usage policies, and, more recently, provided a venue to sign up for notifications on policy updates.

Intended Use Cases From GPT-1 to GPT-4, transparency for models' intended use cases increased with the introduction of model cards and usage policies, bringing more comprehensive and accessible information on models' intended and unintended usage in response to the growing diversity of their application scenarios.

Given its limited capabilities, GPT-1 did not include any guidelines on model use. For GPT-2, OpenAI adopted the newly proposed model card format for reporting, where they specified the primary intended users as AI researchers and practitioners. The model card also outlined unintended uses, such as applications requiring factual output or direct interactions with humans, given GPT-2's limited reliability and prevalence of biases. This model reporting approach facilitated transparent communication of GPT models' intended application.

GPT-3 model card expanded the intended use to broader commercial cases, while vaguely defining prohibited uses as those that "cause societal harms." (OpenAI 2020) According to the usage policies changelog and an examination of available prior versions, throughout the GPT-3 period and the beginning of the GPT-4 period, OpenAI modified usage policies 7 times to detail case-specific requirements and restructure the policy document for better accessibility. For instance, they simplified content and use case policies into one comprehensive guide and changed the visual display of disallowed use cases from interactive to non-interactive. Ten months after GPT-4's release, OpenAI generalized disallowed use cases into four universal rules to further improve accessibility and flexibility.

Monitoring and Regulations Transparency for monitoring and regulating model usage has increased as GPT models progressed, revealing an increasing reliance on automated systems. While OpenAI expressed initial concerns about the challenges of monitoring GPT-2, they have since emphasized strengthing monitoring and regulations as complementary safety measures to model-level mitigations for their subsequent GPT-3 and GPT-4 models.

For GPT-2, OpenAI focused on monitoring public forums for potential malicious use, but did not disclose plans for monitoring the model's actual real-world applications. This monitoring strategy failed to capture the full range of potential misuse, as malicious actors may exploit the model without public announcement. Besides, this approach did not safeguard against the threat of unintentional misuse. For example, deploying GPT-2 in high-stakes decision-making systems could lead to unintended but serious consequences.

The GPT-3 model card first outlined OpenAI's multi-step review process for downstream regulations, from API onboarding to ongoing monitoring of deployed applications for policy violations. According to the recorded changes in usage policies, however, OpenAI largely simplified their monitoring and regulation processes by removing the need for pre-approval (2022-10-25) and then registration (2022-11-09) for creating LLM-infused applications, instead adopting an "outcome-based" monitoring approach that combines "automated and manual methods."

OpenAI provided more information about this half-automated, half-human-reviewed regulation pipeline in their GPT-4 system card, such as the use of content moderation classifiers as an automated tool for monitoring and enforcing usage policies. However, details about the content classifiers, such as their decision-making logic and accuracy, remained unclear. Additionally, the extent and workflow of human oversight, including the specific guidelines and processes of human review within the monitoring pipeline, have not been disclosed. While the shift towards automation may improve efficiency in customer services, greater transparency regarding the current monitoring mechanism is needed to fully evaluate its effectiveness in preventing unintended applications.

Impact Assessment There was a notable increase in the breadth and depth of societal harm assessments and downstream interventions in GPT model reports.

Regarding harm assessment, OpenAI determined GPT-2 was far from usable in practical applications (Radford et al. 2019). Backed on experiments, they anticipated potential harms like model misuse and social biases if deployed, though no real-world harms were observed during the staged release (Solaiman et al. 2019). For GPT-3, OpenAI included a 5-page (approximately 11.9% of the model report) discussion of "broader impacts." They considered broader societal implications, including energy consumption (Brown et al. 2020), deliberate disinformation, and labor market impacts (Tamkin et al. 2021). With a 17-page (approximately 36.2% of the model report) overview of 12 safety challenges associated with model deployment, GPT-4 had the most comprehensive assessment so far, covering inappropriate content (e.g., hallucinations, harmful content), misuse (e.g., privacy, cybersecurity), and other aspects of interoperability with human society (e.g., overreliance, AI acceleration). While expanded assessments showed OpenAI's evolving social awareness, they were mostly theoretical discussions based on model properties and experiments, with limited real-world evidence from actual deployed models.

For deployment-level interventions, OpenAI's approach evolved from researching synthetic text generation for GPT-2, to providing safety guidance to downstream developers for GPT-3, to sharing concrete mitigation tools like content moderation API for GPT-4. This progression suggests OpenAI's growing acknowledgment of the need to collaborate with the wider LLM ecosystem for harm prevention. However, similar to the absence of real-world usage statistics in their impact assessment, OpenAI did not provide information on the efficacy of mitigation tools or examples of their use in practice.

Discussion

According to the findings outlined in the previous section, transparency improvements have mainly focused on application-oriented areas, such as model use cases and impacts, to help assess GPT models' limitations and risks in deployment. Conversely, little to no progress, and even a decline, was witnessed in other key areas related to model development, release, and explainability. Seemingly intending to compensate for the less transparent model-building process, OpenAI provides more information about the individuals and organizations involved in model development to establish liability and accountability.

While these dynamic changes to some extent reflect the dilemma between safety and transparency, as Liang et al. (2022) suggested, there are diverse model release options worth exploring to enhance transparency, rather than forcing a binary decision between full openness and complete secrecy. It remains debatable whether safety concerns sufficiently justify the clear-cut reduction in model development information from GPT-4 onwards. For the rest of the section, we will hypothesize some of the potential causes and implications of these changes.

Potential Causes

The observed transparency changes were likely driven by organizational changes, technological advancements, and societal pressure. From an organizational perspective, OpenAI was subject to increased commercialization after the shift from a nonprofit to a for-profit corporate structure. This shift potentially motivated a reduction in governance transparency to prioritize intellectual property protection and maintain competitive advantages. However, the financial and sociopolitical incentives for GPT's development and deployment were not thoroughly addressed in model documentation.

From a technical perspective, the increased complexity and capability of GPT models have led to their wide adoption for real-world applications, which allowed OpenAI to identify emerging risks and unintended societal impacts, thus contributing to model and application transparency. Nevertheless, technical advancements also aggravated the explainability challenge, which is particularly concerning given that the "black-boxed" GPT models are a crucial component underpinning current safety measures across model-level mitigations, monitoring and regulation mechanisms, and deployment-level interventions (OpenAI 2023).

From a societal perspective, growing public concerns over LLMs' potential impacts might have pressured OpenAI to

improve transparency, particularly application transparency, for stakeholder scrutiny. In the GPT-4 system card, OpenAI cited three of the earliest and most influential scholarly articles (Bender et al. 2021; Weidinger et al. 2022; Ganguli et al. 2022) to introduce some of the identified risks, including bias amplification, ideological cementation, and risky emergent behaviors. However, these articles were only briefly quoted to contextualize OpenAI's own risk taxonomy (OpenAI 2023). This limited engagement with scholarly reviews highlights areas where OpenAI could improve transparent collaboration with external stakeholders to further address potential societal harm.

Implications for LLM Ecosystem

These transparency changes could affect various stakeholders in making informed decisions. For policymakers and regulators, decreased governance transparency could hinder oversight and regulation of the development process, while increased model and application transparency could inform policy decisions and auditing practices to prevent real-world harm. For researchers and developers, decreased governance transparency could limit their abilities to evaluate model risks and collaborate on safety measures, while increased model and application transparency could guide research directions and design choices to promote responsible model use and mitigate unintended consequences. Finally, decreased governance transparency could pose challenges for the general public to hold the model provider accountable for their development and deployment decisions, while increased model and application transparency could help end-users and affected demographics set appropriate expectations for model behaviors and impacts to safely navigate the age of LLMs.

Overall, the implications of transparency evolutions are complex and multifaceted. While the general improvement in model and application transparency could lead to a more responsible LLM ecosystem, missing aspects in these domains could restrict stakeholders' ability to make more informed decisions. For instance, the lack of comprehensive evaluation of model explainability and risk mitigation efficacy could hinder their understanding and prediction of the GPT models' potential harm in real-world applications. Future growth in transparency will be crucial for fostering engagement and collaboration among a broader range of stakeholders to maximize the benefits of this technology while minimizing the risks associated with model use.

Limitations

This study has a limited scope of data selection, as it did not include the full extent of information available about the models, such as OpenAI's blog posts, research publications, and legal documentation. The study also did not trace the evolution of transparency in the intermediary models released between GPT-3 and GPT-4, such as GPT-3.5, which may provide supporting information to interpret the sudden drop in transparency regarding the development process.

Besides, the design of the transparency framework and the evaluation criteria for qualitative analysis were inherently subjective, potentially influenced by the researchers' biases and interpretations. The proposed transparency framework reflected the researchers' value judgments about which aspects of transparency were considered important and desirable. The flexibility in evaluation criteria allowed for subjective interpretations of the disclosed information, such as determining what constituted a discussion of model limitations and the adequacy of that discussion.

Conclusions & Future Work

GPT model reports and usage policies indicate a decline in governance transparency, as marked by a withdrawal of technical details regarding model development, but an increase in model and application transparency over time. The transition to more comprehensive and accessible documentation on model limitations and potential misuse demonstrates the model provider's growing awareness of model weaknesses and societal impacts, leading to more proactive risk mitigation efforts at both the model and application levels. Nevertheless, certain aspects of transparency remain missing, such as model explainability and real-world evidence of LLM impacts. Furthermore, the model reports lack in-depth details on the implementations of safety measures and regulation pipelines.

Future Work

The findings on the evolution of transparency practices for GPT models suggest several avenues for future work. First, future research can empirically investigate the causes and implications of the observed transparency changes, engaging stakeholders to understand their perspectives on effective information disclosure for a more informed transparency framework. Second, future studies could expand the analysis to include transparency of LLMs from other major developers like Anthropic, Google, and Meta to contextualize GPT models' transparency changes within the broader competitive landscape. Longitudinal studies across organizations could reveal industry-wide trends in motivations and challenges around transparency practices. Lastly, research into best practices for transparency is essential to establish evidence-based norms for responsible development and deployment of LLMs. Studying diverse transparency approaches may lead to innovative solutions for balancing safety and openness.

Acknowledgments

We would like to thank Professor Julie Walsh and other Laboratory for Ethics, Equity, and Digital Technology (LEED) members at Wellesley College. We also acknowledge funding from the NSF ER2 2220772 grant.

References

Arrieta-Ibarra, I.; Goff, L.; Jiménez-Hernández, D.; Lanier, J.; and Weyl, E. G. 2018. Should we treat data as labor? Moving beyond "free". In *AEA Papers and Proceedings*, volume 108, 38–42. American Economic Association 2014 Broadway, Suite 305, Nashville, TN 37203.

Ball, C. 2009. What is transparency? *Public Integrity*, 11(4): 293–308.

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.

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:2108.07258v1.

Bommasani, R.; Klyman, K.; Longpre, S.; Kapoor, S.; Maslej, N.; Xiong, B.; Zhang, D.; and Liang, P. 2023. The Foundation Model Transparency Index. arXiv:2310.12941.

Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J. D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33: 1877–1901.

Chen, L.; Zaharia, M.; and Zou, J. 2023. How is ChatGPT's behavior changing over time? arXiv:2307.09009.

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

Du, N.; Huang, Y.; Dai, A. M.; Tong, S.; Lepikhin, D.; Xu, Y.; Krikun, M.; Zhou, Y.; Yu, A. W.; Firat, O.; Zoph, B.; Fedus, L.; Bosma, M.; Zhou, Z.; Wang, T.; Wang, Y. E.; Webster, K.; Pellat, M.; Robinson, K.; Meier-Hellstern, K.; Duke, T.; Dixon, L.; Zhang, K.; Le, Q. V.; Wu, Y.; Chen, Z.; and Cui, C. 2022. GLaM: Efficient Scaling of Language Models with Mixture-of-Experts. arXiv:2112.06905.

Ferrara, E. 2023. Should chatgpt be biased? challenges and risks of bias in large language models. arXiv:2304.03738.

Ganguli, D.; Hernandez, D.; Lovitt, L.; Askell, A.; Bai, Y.; Chen, A.; Conerly, T.; Dassarma, N.; Drain, D.; Elhage, N.; et al. 2022. Predictability and surprise in large generative models. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, 1747–1764.

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.

Glaser, V. L.; Pollock, N.; and D'Adderio, L. 2021. The biography of an algorithm: Performing algorithmic technologies in organizations. *Organization Theory*, 2(2): 1–27.

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

Hacker, P.; Engel, A.; and Mauer, M. 2023. Regulating ChatGPT and other large generative AI models. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, 1112–1123.

Helmond, A.; Nieborg, D. B.; and van der Vlist, F. N. 2019. Facebook's evolution: Development of a platform-as-infrastructure. *Internet Histories*, 3(2): 123–146.

Higgins, C.; and Tang, S. 2024. Organisational integrity and transparency. In *Research Handbook on Organisational Integrity*, 470–484. Edward Elgar Publishing.

Hrín, A. 2023. *Methods for investigating the external and internal validity of machine learned signals*. Master's thesis, University of Helsinki.

Hyysalo, S.; Pollock, N.; and Williams, R. 2019. Method matters in the social study of technology: Investigating the biographies of artifacts and practices. *Science & Technology Studies*, 32(3): 2–25.

Internet Archive. 2001. Wayback Machine. https://web.archive.org. Accessed: 2024-03-21.

Jobin, A.; Ienca, M.; and Vayena, E. 2019. The global landscape of AI ethics guidelines. *Nature machine intelligence*, 1(9): 389–399.

Kaplan, J.; McCandlish, S.; Henighan, T.; Brown, T. B.; Chess, B.; Child, R.; Gray, S.; Radford, A.; Wu, J.; and Amodei, D. 2020. Scaling Laws for Neural Language Models. arXiv:2001.08361.

Lacoste, A.; Luccioni, A.; Schmidt, V.; and Dandres, T. 2019. Quantifying the Carbon Emissions of Machine Learning. arXiv:1910.09700.

Li, H.; Vincent, N.; Chancellor, S.; and Hecht, B. 2023. The dimensions of data labor: A road map for researchers, activists, and policymakers to empower data producers. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, 1151–1161.

Liang, P.; Bommasani, R.; Creel, K.; and Reich, R. 2022. The time is now to develop community norms for the release of foundation models. https://hai.stanford.edu/news/time-now-develop-community-norms-release-foundation-models. Accessed: 2024-03-21.

Liao, Q. V.; and Vaughan, J. W. 2023. AI Transparency in the Age of LLMs: A Human-Centered Research Roadmap. arXiv:2306.01941.

Liu, Z.; Qiao, A.; Neiswanger, W.; Wang, H.; Tan, B.; Tao, T.; Li, J.; Wang, Y.; Sun, S.; Pangarkar, O.; Fan, R.; Gu, Y.; Miller, V.; Zhuang, Y.; He, G.; Li, H.; Koto, F.; Tang, L.; Ranjan, N.; Shen, Z.; Ren, X.; Iriondo, R.; Mu, C.; Hu, Z.;

Schulze, M.; Nakov, P.; Baldwin, T.; and Xing, E. P. 2023. LLM360: Towards Fully Transparent Open-Source LLMs. arXiv:2312.06550.

Luccioni, A. S.; and Hernandez-Garcia, A. 2023. Counting Carbon: A Survey of Factors Influencing the Emissions of Machine Learning. arXiv:2302.08476.

Mallin, C. 2003. The relationship between corporate governance, transparency and financial disclosure. *Transparency and Financial Disclosure*.

Mantelero, A. 2018. AI and Big Data: A blueprint for a human rights, social and ethical impact assessment. *Computer Law & Security Review*, 34(4): 754–772.

Masotina, M.; Musi, E.; and Spagnolli, A. 2023. Transparency is Crucial for User-Centered AI, or is it? How this Notion Manifests in the UK Press Coverage of GPT. In *Proceedings of the 15th Biannual Conference of the Italian SIGCHI Chapter*, 1–8.

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.

Mökander, J.; Schuett, J.; Kirk, H. R.; and Floridi, L. 2023. Auditing large language models: a three-layered approach. *AI and Ethics*, 1–31.

Narayanan, A.; and Kapoor, S. 2023. Generative AI companies must publish transparency reports. http://knightcolumbia.org/blog/generative-ai-companies-must-publish-transparency-reports. Accessed: 2024-03-21.

OpenAI. 2019. GPT-2 Model Card. https://github.com/openai/gpt-2/blob/master/model_card.md. Accessed: 2024-03-21.

OpenAI. 2020. GPT-3 Model Card. https://github.com/openai/gpt-3/blob/master/model-card.md. Accessed: 2024-03-21.

OpenAI. 2023. GPT-4 Technical Report. arXiv:2303.08774v4.

OpenAI. 2024. Usage Policies. https://openai.com/policies/usage-policies. Accessed: 2024-03-21.

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:2104.10350.

Perrigo, B. 2023. Exclusive: OpenAI used Kenyan workers on less than $2 per hour to make ChatGPT less toxic. *Time Magazine*, 18.

Radford, A.; Narasimhan, K.; Salimans, T.; and Sutskever, I. 2018. Improving language understanding by generative pre-training. https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf. Accessed: 2024-03-21.

Radford, A.; Wu, J.; Child, R.; Luan, D.; Amodei, D.; and Sutskever, I. 2019. Language models are unsupervised multitask learners. https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf. Accessed: 2024-03-21.

Reisman, D.; Schultz, J.; Crawford, K.; and Whittaker, M. 2018. Algorithmic Impact Assessments Report: A Practical Framework for Public Agency Accountability. *AI Now Institute*, 9.

Rudin, C. 2019. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature machine intelligence*, 1(5): 206–215.

Schiff, D.; Rakova, B.; Ayesh, A.; Fanti, A.; and Lennon, M. 2020. Principles to Practices for Responsible AI: Closing the Gap. arXiv:2006.04707.

Schwartz, R.; Dodge, J.; Smith, N. A.; and Etzioni, O. 2020. Green AI. *Communications of the ACM*, 63(12): 54–63.

Selbst, A. D. 2021. An institutional view of algorithmic impact assessments. *Harv. JL & Tech.*, 35: 117.

Selbst, A. D.; Boyd, D.; Friedler, S. A.; Venkatasubramanian, S.; and Vertesi, J. 2019. Fairness and abstraction in sociotechnical systems. In *Proceedings of the conference on fairness, accountability, and transparency*, 59–68.

Singh, C.; Inala, J. P.; Galley, M.; Caruana, R.; and Gao, J. 2024. Rethinking Interpretability in the Era of Large Language Models. arXiv:2402.01761.

Solaiman, I.; Brundage, M.; Clark, J.; Askell, A.; Herbert-Voss, A.; Wu, J.; Radford, A.; Krueger, G.; Kim, J. W.; Kreps, S.; McCain, M.; Newhouse, A.; Blazakis, J.; McGuffie, K.; and Wang, J. 2019. Release Strategies and the Social Impacts of Language Models. arXiv:1908.09203.

Solaiman, I.; Talat, Z.; Agnew, W.; Ahmad, L.; Baker, D.; Blodgett, S. L.; au2, H. D. I.; Dodge, J.; Evans, E.; Hooker, S.; Jernite, Y.; Luccioni, A. S.; Lusoli, A.; Mitchell, M.; Newman, J.; Png, M.-T.; Strait, A.; and Vassilev, A. 2023. Evaluating the Social Impact of Generative AI Systems in Systems and Society. arXiv:2306.05949.

Strubell, E.; Ganesh, A.; and McCallum, A. 2019. Energy and Policy Considerations for Deep Learning in NLP. arXiv:1906.02243.

Tamkin, A.; Brundage, M.; Clark, J.; and Ganguli, D. 2021. Understanding the Capabilities, Limitations, and Societal Impact of Large Language Models. arXiv:2102.02503.

Turilli, M.; and Floridi, L. 2009. The ethics of information transparency. *Ethics and Information Technology*, 11: 105–112.

Weidinger, L.; Uesato, J.; Rauh, M.; Griffin, C.; Huang, P.-S.; Mellor, J.; Glaese, A.; Cheng, M.; Balle, B.; Kasirzadeh, A.; et al. 2022. Taxonomy of risks posed by language models. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, 214–229.

Wiegel, V. 2016. *Biographies of an innovation: an ecological analysis of a strategic technology project in the auto-industry*. Phd thesis, The University of Edinburgh.

Zhang, Y.; Li, Y.; Cui, L.; Cai, D.; Liu, L.; Fu, T.; Huang, X.; Zhao, E.; Zhang, Y.; Chen, Y.; Wang, L.; Luu, A. T.; Bi, W.; Shi, F.; and Shi, S. 2023. Siren's Song in the AI Ocean: A Survey on Hallucination in Large Language Models. arXiv:2309.01219.