# Treading the Transparency Tightrope: A Taxonomy of Risks and Benefits of Foundation Model Data Transparency for Transparency Advocates

Morgan Klaus Scheuerman
AI Ethics
Sony AI
Broomfield, Colorado, USA
morgan.scheuerman@sony.com

Wiebke Hutiri
AI Ethics
Sony AI
Zurich, Switzerland
wiebke.hutiri@sony.com

Aida Rahmattalabi
AI Ethics
Sony AI
Los Angeles, California, USA
aida.rahmattalabi@sony.com

Victoria Matthews
AI Ethics
Sony AI
New York, New York, USA
victoria.matthews@sony.com

Alice Xiang
AI Ethics
Sony AI
Seattle, Washington, USA
alice.xiang@sony.com

Jerone Andrews
AI Ethics
Sony AI
London, United Kingdom
jerone.andrews@sony.com

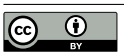
## Abstract

Data powering AI is often opaque. Researchers, NGOs, and law and policy leaders have called for greater transparency about how data is used for training, fine-tuning, and evaluation. While data transparency is often championed as crucial, what it concretely enables is largely implicit. Similarly, the concerns developers seem to have about transparency go unstated. This lack of clarity has led some researchers to critique transparency demands as disconnected from the actual benefits—or risks—to specific stakeholders. We analyze documentation from four stakeholder groups to create a taxonomy of the risks and benefits of dataset transparency. Data transparency is perceived as either a risk or a benefit given a stakeholder's position, rather than wholesale. We also propose data availability and data documentation as two lenses through which to consider transparency. We discuss how best to strategically promote *situational data transparency* that takes into account the relationship between stakeholder position, transparency modality, and benefits/risks.

## Keywords

Foundation models, data transparency, datasets, generative AI, privacy, data provenance, open-source, technology companies

**ACM Reference Format:**
Morgan Klaus Scheuerman, Wiebke Hutiri, Aida Rahmattalabi, Victoria Matthews, Alice Xiang, and Jerone Andrews. 2026. Treading the Transparency Tightrope: A Taxonomy of Risks and Benefits of Foundation Model Data Transparency for Transparency Advocates. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems (CHI '26), April 13–17, 2026, Barcelona, Spain.* ACM, New York, NY, USA, 29 pages. https://doi.org/10.1145/3772318.3790364




*CHI '26, Barcelona, Spain*

ACM ISBN 979-8-4007-2278-3/26/04
https://doi.org/10.1145/3772318.3790364

## 1 Introduction

Large AI systems trained on extensive data, known as *foundation models* (FMs) [22, 45, 113], have become central to recent advancements in AI due to their versatility across a wide range of applications [51, 59, 186, 202]. However, the developers of these models face criticism over how they collect and use *data* (e.g., [42, 95, 134, 135, 211]). Unlike task-specific narrow models, such as face detectors and recommender systems, FMs require vastly more data for training, fine-tuning, and evaluation [112, 219, 243], leading to intense scrutiny of data scaling practices.

FM developers often scrape data from the web [73, 117]. Many have highlighted harms that result from this practice, raising concerns about privacy breaches, lack of consent, harmful content, intellectual property violations (e.g., [24, 26, 135])—and how this data influences model behavior and outcomes [42, 154, 211]. Little is known about the datasets used to pre- and post-train the most profitable and ubiquitous FMs on the market [47, 88, 235]. Commercial developers, in particular, often limit documentation, making insights into these datasets inaccessible to external stakeholders. Researchers and law and policy leaders have expressed concerns that the opacity of commercial AI datasets impedes independent analysis of the harms of black-box proprietary models [60, 95, 169, 189], even as these models become increasingly integrated into daily life. Some have critiqued the current trajectory of AI data mining practices for failing to account for social good goals, which they believe would be better enabled by transparency [75, 153, 221, 224].

*Data transparency* has been proposed as a solution to address the concerns associated with AI datasets, broadly, and FM datasets, specifically. The implicit argument is that greater visibility into how datasets are constructed and what they contain could help mitigate associated concerns [154]. However, as critical transparency scholars have argued far before the advent of FMs, transparency is not inherently good, but a principle that can enable both benefits *and* risks [69, 227], contingent on context [82, 177]. Thus, to position AI data transparency as a precondition for affected stakeholders to meaningfully engage in the current AI data ecosystem, we need to better understand the perceived benefits of transparency, as well as

the perceived risks. As Corbett and Denton argue about how transparency has been approached in responsible AI, "We need to start framing the value of transparency in a contingent manner instead of the current norm of presuming transparency has an inherent normative value" [66].

In this work, we build on Corbett and Denton's call to better understand the underlying incentives of AI data transparency, specifically in the current AI data ecosystem dominated by FMs. By framing transparency as contingent, we thus argue that it is necessary to approach transparency as a value that is rooted in specific stakeholder needs and perspectives [83]. We examine how AI data transparency can pose *risks* and present *benefits* by analyzing documents from four different stakeholder groups in the AI transparency space: researchers, NGOs, law and policy leaders, and commercial and community developers [78]. In the context of this paper, we specifically situate our use of the term "transparency" to mean AI data transparency, broadly characterized as *practices, approaches, and demands associated with making visible and accessible some, or all, characteristics of the data underlying training, fine-tuning, and evaluating models, including but not limited to the processes underlying data curation, the provenance of the data, the composition of the data, and the (raw and/or processed) data itself* [38, 39, 55, 165, 235].

Our key contribution is a *taxonomy of risks and benefits* of AI data transparency. We surface four perceived risks (contamination, competitiveness, safety, and scrutiny) and four perceived benefits (accountability, innovation, integrity, and suitability) of AI data transparency. We also propose two factors influential to transparency perceptions. The first factor is *stakeholder position*: whether a stakeholder takes a stance of transparency advocate, pushing for data transparency, or opposition, seeking to conceal aspects of AI data. The second factor we identify is *transparency modality*, either in the form of *data availability* or *data documentation*. We argue that the degree of transparency for either modality impacts the range of risks and benefits that may impact relevant stakeholders.

We take the position that the benefits of transparency should be prioritized, particularly because many of the data transparency risks primarily affect those who stand to gain from opaque data practices—even if they harm other stakeholders. Our taxonomy is aimed at guiding transparency advocates towards *situational data transparency*: taking into account the benefits and risks transparency may present to specific stakeholders with specific goals using specific modalities. To be more effective in their advocacy, advocates should align situational data transparency interventions with specific goals, such as accountability to legal frameworks, rather than broad calls that position transparency as an assumed inherent good. For example, transparency advocates may choose to articulate goals in ways that convince developers to adopt transparency within their existing incentive mechanisms or dispute otherwise unfounded arguments for data opacity. We hope that our taxonomy aids data transparency advocates in surfacing and executing benefits while navigating the risks. As such, we conclude this paper with a range of considerations for promoting situational data transparency more effectively. For HCI researchers working on responsible AI data practices, we discuss how our taxonomy can act as a grounding framework for future research.

## 2 Related Work: Opposing Opacity

We summarize literature on AI data transparency practices and approaches. We then describe critiques of how transparency has been mainstreamed in responsible AI. We describe how approaches have been divorced from *critical transparency scholarship*, resulting in an approach to dataset transparency that prioritizes explainability but fails to account for efficacy or impact. Critical transparency scholarship informs our approach to developing the taxonomy presented in this paper.

### 2.1 Transparency: The Antidote to AI Harms?

Transparency is a foundational principle proposed to mitigate many of the major issues associated with AI [3]. While transparency is, in reality, a nebulous concept with numerous definitions [23, 225], data transparency can be broadly defined as the ability to perceive or gain access to information about the underlying data within a system, including characteristics like collection processes, composition, and provenance [39]. Transparency can be thought of in terms of *visibility*; "the possibility of accessing information, intentions or behaviours ... revealed through a process of disclosure" [214].

Generally, the data used to train, fine-tune, and evaluate FMs are not released to the public. Often, FM datasets are so minimally documented that their composition is entirely unknown. This opacity is not a new phenomenon; it has existed since companies commercialized narrow AI models (e.g., [138, 150, 183, 201, 216]). However, the scale of FMs has brought data transparency concerns even more to the forefront than for their narrow model counterparts [31, 136, 219]. Recently, Pepe et al. found that only 14% of pretrained transformer-based machine learning models on Hugging Face specify their training datasets in model cards [162]. Without deeper understanding of the data used to train these models, it is difficult for stakeholders to anticipate model limitations, yet alone mitigate harms.

While academic researchers still tend to embrace transparency in the form of open-source datasets [122, 178, 232], commercial developers rarely release the datasets used in their models [2, 56, 149, 173]. There has been ongoing pressure on companies developing AI to be more transparent about their data practices and use. The logic behind these calls is, in part, as follows: being more transparent enables people to do something in response to the potential or real harms surfaced by AI. Worth et al. argue that "AI *data* transparency is needed to address the particular limitations for public accountability of AI systems" (emphasis in original) [235]. Scholars have developed databases, measures, and trackers for assessing the overall transparency of FMs. For example, the Foundation Model Transparency Index rates 14 FMs on a transparency scale from 0-100, with one dimension including dataset transparency [47]. Others have developed mechanisms for documenting data curation choices. Some of the most known transparency documentation frameworks include Datasheets for Datasets [84], Healthsheets [175], and Dataset Nutrition Labels [102].

Alongside pushes for transparency are drives for more open-source AI models [7, 30, 35, 99, 140, 231], though the definition of open-source AI remains somewhat in contention. In 2024, the Open Source Initiative (OSI), a non-profit organization focused on promoting the open-source software movement, released an

official definition [172]. For an AI model to be open-source by this definition, the developer must include: "1) the complete description of all data used for training, including (if used) of unshareable data, disclosing the provenance of the data, its scope and characteristics, how the data was obtained and selected, the labeling procedures, and data processing and filtering methodologies; (2) a listing of all publicly available training data and where to obtain it; and (3) a listing of all training data obtainable from third parties and where to obtain it, including for [free]" [11]. Commercial models previously branded as open-source, like Llama and Gemma, fail to meet criteria for this definition due to data opacity.

Data transparency is also emerging as key within AI legislation, with various data documentation requirements enshrined in laws in, for example, the European Union (EU) [14], United States (US) [9], and China [5]. In the United States, California law AB-2013 similarly requires developers—and even those that fine-tune or re-train models—to provide publicly available information about the training data of certain generative AI systems [1]. China's Measures for Generative AI requires information about generative AI data sources and labeling to be shared with regulators only if necessary for enforcement, rather than proactively mandating disclosure [2]. The EU stands out for providing a tool: a training data transparency template for providers of certain general-purpose AI models [15].

However, there is a lack of standardization regarding data transparency. Legal approaches to enforcing AI data transparency may diverge in both content and prescriptiveness, leading to calls for greater global standardization and coordination of data documentation artifacts [104]. The current global policy landscape suggests we may see continued theoretical agreement about the importance of data transparency, but different approaches in practice, and at various levels of abstraction ([9, 18].

## 2.2 Critical Transparency: Aligning Transparency with Action

Clearly, many different stakeholders—including researchers, NGOs (e.g., OSI), law and policy leaders, and the developers of open-source models—are concerned about the transparency of AI data. While some developers curate opaque datasets, both "soft" calls (e.g., agreements, policy, definitions) and "hard" calls (i.e., laws and regulations) seek to push the boundaries of AI data from opaque to transparent. However, simply advocating that developers be transparent does not provide clarity on *what* data transparency should achieve. Even when transparency is mandated by law (e.g., [5, 9, 14]), the actions that transparency may enable for stakeholders are often left implicit or vague.

Corbett and Denton interrogated how transparency approaches regularly fail to enable positive or productive impacts on the fairness or accountability goals central to the responsible AI community. They critiqued the assumptions that transparency itself "is an effective means of making AI systems more fair and accountable" and "in general, is a societal good and therefore should be pursued as an end in itself" [66]. Even if some stakeholders might view transparency as an inherent societal good [225], the call for transparency in AI has become a trope that is divorced from specific situational intent. Instead, it has largely been used as a synonym for explainability, purporting that explanation alone is interchangeable with intervention [66]. More direct actions which should be enabled by transparency are rarely measured or evaluated [66]. As Gray et al. argue about dark patterns [90], a lack of clarity about the underlying dimensions defining transparency may limit transdisciplinary efforts to advocate for AI data transparency in an era where commercial developers vie to define what should and should not be transparent [85].

Further, approaches to transparency in responsible AI research have often been divorced from the rich scholarship on critical transparency studies, which appropriately situates transparency as an ambiguous principle that enables or disables certain actions or values. From this perspective, transparency is not inherently positive; it can have many negative implications [23, 69, 125, 127, 225, 227]. Critical transparency scholars have highlighted the potential risks associated with naive calls for transparency, including: enabling surveillance [50, 147, 155], harming vulnerable groups [25, 37, 159], and obfuscating more pertinent regulatory and democratic interventions [23, 41, 127]. As Keyes et al. satirically articulate in [116], providing transparency about an algorithm designed to eliminate elderly people with low social credit scores does not make such an algorithm ethical or moral; it simply allows you to know what morally reprehensible actions are being taken by the algorithm. Or as Xu and Mustafaraj summarize: "transparency alone does not guarantee ethical outcomes" [239].

In reality, transparency is neither an inherent benefit or risk. Rather, transparency shifts information asymmetries towards or away specific stakeholders, enabling them to take specific actions [79, 115, 199]. Transparency is thus a *value* that the designers of datasets may choose, or choose not, to incorporate [83, 176, 247]. Thus portrayed, how data transparency is, or is not, presented enables strategic behaviors, which may result in both benefits and harms. Corbett and Denton proposed ways that responsible AI scholars can steer transparency demands towards more meaningful and productive impacts: by reclaiming transparency from explainability, making transparency contingent, and centering how transparency mechanisms impact people [66].

In this work, we employ a longstanding HCI tradition of centering stakeholder values [83, 176, 247] examine the contingent reasons different affected stakeholder types may advocate or oppose AI data transparency. Centering different stakeholder perspectives can raise different values, which is crucial when considering what critical transparency might enable or disable for those stakeholders. We assess whether these perceived risks and benefits are associated with specific *modalities* of dataset transparency. As such, we build Corbett and Denton's invitation to better situate transparency within the kinds of results that it can actually enable, beneficial or otherwise. We can then argue more strategically for the underlying benefits that data transparency should enable, given its specific modality, and to more concretely mitigate associated risks. Through providing a taxonomy of risks and benefits—taking into account stakeholder positions and transparency modalities—we also extend recent literature on better defining properties of responsible AI (e.g., [160]).

[1] see §1(b-d) of [9]
[2] see Article 19 of [5]

# 3 Method: Towards Making Dataset Transparency Contingent

Inspired by prior work focused on conceptually classifying the impacts of algorithmic technologies via engagement with prior literature (e.g., [124, 180]), like FMs (e.g., [74]), our goal is to create a conceptual taxonomy [32] of AI data transparency through a ground-up empirical analysis of existing literature on AI data transparency. We now describe our approach to defining the contingent risks and benefits of AI dataset transparency.

## 3.1 Sampling scope.

We sample documentation using different actors in the AI transparency space. In this work, we have chosen to simplify the number of stakeholders that we analyzed for several reasons. We sought primarily to understand how transparency might impact stakeholders with specific roles in the FM ecosystem—development, research, regulation, and advocacy. These four roles were derived from a clustering of actions associated with Eyert and Lopez's eight actors (e.g., machine learning practitioners, companies, and some scholars and NGOs participate in development; researchers and scholars, NGOs, politicians, activists, journalists, and political decision makers participate in advocacy) [78]. We note that some stakeholder perspectives overlapped. For example, when sampling for researchers, we also identified academic papers focused on the development of transparent models (e.g., [119, 133]). **We sampled literature and documentation from: (1) academic researchers, (2) NGOs, (3) law and policy leaders, and (4) commercial and community model developers.**

*3.1.1 Sampling limitations.* We acknowledge that these stakeholder groups do not cover the full spectrum of those invested in or impacted by AI or FMs. It is also crucial to note that many stakeholders (such as community developers) have discussions about transparency outside of academic publication, making documents and discussions more difficult to systematically sample. For example, while the authors of Llama [89] and Gemini [204] discuss efforts to minimize safety risks, they do not openly discuss other reasons for limiting disclosure to their data, either in the cited white papers or in blog posts. Data subjects are also a crucial stakeholder group in AI data transparency discussions, but often largely represented via academic researchers focused on human subjects research. Similar to community developers, discussions of transparency directly from known or potential data subjects may occur on more informal sources like forums. We openly acknowledge that certain perspectives may be underrepresented in our analysis, given they did not surface in the systematic sampling approaches we developed. Furthermore, it was more straightforward to systematically sample works from researchers and developers, given how such works are indexed compared to those of NGOs and law and policy leaders, especially given the rapidly evolving legal and policy landscape. For this reason, we purposively sampled from NGOs and law and policy (see Section 3.2.2 and Section 3.2.3), which we acknowledge means that these stakeholders are underrepresented in comparison to researchers and developers, in particular. Future work specifically focused on these underrepresented stakeholder groups would be of value.

We also note that, while we explicitly chose to include China's Interim Measures for the Management of Generative Artificial Intelligence Services [5], our sample largely represents Anglophone countries. As prior scholars have stated, many computing venues, like CHI and FAccT, remain dominated by WEIRD nations [132, 185]. However, this limitation is also due to our language proficiency being English, leading us to conduct sampling in English. Future work on data transparency would benefit from explicitly non-English sources.

Overall, documents reflecting different stakeholder perspectives on FM transparency are in a space of rapid development (e.g., policy and legal frameworks). We sought to represent diverse stakeholder perspectives without seeking to attempt to ensure our corpus was whole and complete. However, the documents they produce present an opportunistic starting point for understanding the perceived risks and benefits of data transparency. We did not weigh certain stakeholder perspectives as more important in defining risks and benefits, but sought to identify trends that arose across and between stakeholders. We note numerous areas for future work given these limitations in Section 7.

## 3.2 Document Selection

We examined 153 sources. We sought to gather a diverse set of literature to ensure that we represent the perspectives of four different types of stakeholders involved in AI dataset transparency discussions: researchers, NGOs, law and policy leaders, and developers. A modified PRISMA Flow Diagram [156] describing our document selection process across stakeholders can be found in Appendix A.1.1. A corpus of our sources can also be found in Appendix A.1.2. Finally, our coding materials can be found at https://doi.org/10.5281/zenodo.18340221.

*3.2.1 Researchers.* We examined 91 academic research papers. Given the nature of academic publication to be indexed by discipline and topic, we took a systematic sampling approach to our selections. We first targeted academic work on data transparency by targeting two scholarly communities in the ACM: FAccT and CHI. We chose to target work published at FAccT because it is: (1) explicitly dedicated to scholarship focused on fairness, accountability, and—most crucially—transparency in algorithmic systems; and (2) as a transdisciplinary venue, the role that HCI scholarship is playing continues to increase. Meanwhile, we chose to include works at CHI not only because it remains the premier HCI venue, but because human-centered AI (HCAI) topics like data transparency continue to grow in its proceedings. Further, the literature at CHI often differs from the literature at FAccT. While FAccT works often presented insights based on literature and case studies, CHI works often included talking directly with stakeholders, including evaluating data transparency prototypes. While the focus of works at CHI were not necessarily the transparent sharing of data used to train models like works seen at FAccT, user-centric data transparency still provided insights into the potential risks and benefits of data transparency. We acknowledge that other communities, such as AIES and AAAI, are also scholarly communities with work on transparency in AI systems. We chose to focus on FAccT and CHI in this paper, specifically, because both remain the premiere conferences for both transparency in AI and HCI, respectively.

Finally, we decided to augment our research by exploring transdisciplinary scholarship indexed on Google Scholar. We chose to use Google Scholar because we sought perspectives from researchers outside of ACM publications, as well as works in progress, as published on arXiV. Ultimately, our goal was not to comprehensively cover the perspectives of either HCI or responsible AI scholars, but to examine a range of differing perspectives. We acknowledge the limitations of Google Scholar for sourcing literature, particularly in its lack of reproducibility.

**FAccT sampling.** To source literature from FAccT, we searched the FAccT proceedings in the ACM Digital Library for papers with "dataset transparency," "data sharing," or "open data" in their titles, abstract, or author keywords. This yielded 92 papers. However, this number was misleading, both because of the colloquial use of transparency in "FAccT" and because papers did not always explicitly focus on issues of data transparency. We reviewed each paper to ensure that it focused on data transparency. In total, we selected 34 papers on data transparency.

**CHI sampling.** When we used similar keywords as our approach to FAccT sampling to source literature specific to AI data transparency from CHI, we found that, despite an overwhelming number of results (2,468), most papers had nothing to do with data transparency. As such, we decided to focus solely on full papers in the ACM Digital Library for papers with "data transparency" in their titles, abstract, or author keywords. Much as we describe below on including arXiv, we did not remove extended abstracts as we sought to understand any contingent actions tied to data transparency. This approach yielded 114 results. We then reviewed each result to ensure that references to data transparency were to AI or FMs. We did not include papers focused on data transparency for other technical services, like apps that did not clearly use machine learning or AI. In total, we sampled 21 papers.

**Google Scholar sampling.** We augmented our research sample by examining literature outside of FAccT and CHI. We used Google Scholar to identify the most relevant papers with the search term "foundation model dataset transparency." While we removed papers unrelated to dataset transparency (e.g., focused on non-profit foundations), we did not remove papers due to their publication origin (e.g., arXiv) or focus (e.g., critiques, frameworks). We sourced papers until we reached 50, then refined the sample by excluding those not explicitly addressing AI or FM transparency. This yielded 36 papers.

*3.2.2 NGOs.* We examined documentation from three NGOs that focus on AI transparency: (1) Open-Source Initiative (OSI); (2) Partnership on AI (PAI); and (3) The Transparency Coalition. We used purposive sampling [157] to select these three NGOs, because they have openly critiqued commercial AI and FM transparency practices and been involved in developing data transparency definitions and frameworks. We were already familiar with these NGOs given their role in ongoing research and other initiatives in the AI transparency space. We examined statements about transparency on their websites, as well as reports, white papers, and other documentation relevant to their stances on AI data transparency.

*3.2.3 Law and Policy Leaders.* As previously highlighted in Section 2.1, transparency has already been enshrined in AI legislation. Additionally, it underlies many AI policy frameworks. We examined nine frameworks, either already established as critical transparency artifacts of law and policy or clearly indicative of them: (1) Executive Order 14110 on Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence [154][3]; (2) The EU AI Act [14]; (3) Hiroshima Code of Conduct [4]; (4) NIST-AI-600-1, Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile [154]; (5) OECD's Recommendation of the Council on Artificial Intelligence [13]; (6) Seoul Frontier AI Commitments [8]; (7) the UK's "A pro-innovation approach to AI regulation" white paper [12]; (8) Generative Artificial Intelligence: Training Data Transparency, AB-2013, California Legislature (2023–24) [9]; and (9) China's Interim Measures for the Management of Generative Artificial Intelligence Services [5]. We sourced most frameworks from Howell and Ifayemi [104], but also purposively included AB-2013 [9] and the Interim Measures for the Management of Generative Artificial Intelligence Services [5], given the specificity of AI data transparency measures in both.

While these nine frameworks offer baseline approaches to data transparency by legal and policy leaders, their ability to remain evergreen is dependent on global policy changes. It remains to be seen how these frameworks may influence future compliance, governance, standardization, and industry efforts on transparency.[4]

*3.2.4 Developers.* As described in the introduction, the proliferation of FMs are currently driving the AI landscape and opening up new conversations about data transparency. Therefore, we chose to focus on documentation associated with FMs. We examined documentation from 50 FMs. We examined any explicit justifications for or against data transparency, as well as more implicit stances about data transparency (e.g., statements about proprietary data serving the needs of customers insinuating a desire for maintaining a competitive advantage). We included fully closed-source proprietary models (e.g., Google's Gemini, OpenAI's GPT); open-weight models (e.g., Meta's Llama, Google's Gemma); and fully open-source (i.e., the data is freely available) models (e.g, Cohere's Aya, BigScience's BLOOM). We sourced models from the LMSYS Chatbot Arena Leaderboard, citations in research papers (e.g., [47, 198]), and from legal cases (e.g., [33, 53, 197]) until we reached 50 FMs. We eliminated FMs that did not have sufficient documentation to analyze in the form of robust model cards, technical reports, or papers (e.g., XAI's Grok). Given that reporting on training and fine-tuning data in the selected model documentation is still sparse, we also examined websites, data and model cards, and sometimes Hugging Face forums linked to official model cards, when available. Nonetheless, some developers never explicitly state, or even implicitly imply, their stance on data transparency in any of these documents (e.g., Qwen3, Command A, Yi). While we acknowledge that there are many other closed-source, open-weight, and open-source models

[3]While this Executive Order has been revoked as of January 20, 2025 [16], we believe it is still a relevant example highlighting AI transparency in policy by a previous administration.

[4]New and evolving legal obligations are likely to provide updated regulatory guidance on AI data transparency provisions. For example, the passage of South Korea's AI Basic Act, effective January 2026, has created conditions for future potential regulatory guidance and enforcement actions that cover key transparency provisions for high-impact AI. Additionally, the ongoing implementation of provisions in existing laws like the EU AI Act may shed more light on compliance transparency obligations, as will the General-Purpose AI Code of Practice, Guidelines, and Training Data Template. Finally, shifts in political leadership, like administration changes, could always inform future law and policy on AI data transparency.

out there, we did not seek to be fully comprehensive in our list of existing FMs. Instead, we sought diverse perspectives from different developers.

### 3.3 Analysis Approach

To understand the potential risks and benefits of AI data transparency, we analyzed statements surrounding transparency in our corpus. We examined statements that promote and justify AI data transparency, as well as those that either explicitly or implicitly resist transparency. Many of the documents we analyzed discussed the transparency of data explicitly. However, because many documents discuss transparency as relevant to AI or FMs more broadly, we also decided to extrapolate how these value statements could pertain to data more specifically.

The first author performed initial line-by-line open coding [171]. The first author coded for what we perceived as either (1) the risks or (2) the benefits associated with data transparency. Coding was informed by the question: *What does this statement say or insinuate that data transparency enables?* For example, we saw statements that insinuated releasing a dataset may risk leaking personally identifiable information about data subjects. Some statements also included sentiments about both risks and benefits. For instance, the following statement promotes the sharing of AI training data so that external parties can audit that data (and, presumably, its connection to model outputs), but at the same time acknowledges the risks that developers are concerned about, namely competitiveness:

> *"Major AI labs continue treating full-scale training data specifics as proprietary secrets,* ***citing competitive advantages around quality or scale****, but leaving* ***auditing*** *near impossible* ***for those affected by AI services****"* [86].

After coding excerpts based on what transparency enables, the research team met to collectively refine the open codes into more specific lower-level concepts. We then began to define them through examples in the data. Instances could be assigned more than one concept. For instance, the example above was labeled with "privacy" and "competitiveness." During discussions, we also decided to discard codes that were prevalent but not necessarily relevant to our goal of defining contingent risks and benefits. For example, we saw many stakeholders state the benefit of data transparency as trustworthiness." However, as we note further in Section 7, trustworthiness is also a vague concept largely lacking articulations of the specific actions it enables.

Following the phase of defining lower-level coding concepts, the first author then began to thematically cluster each code into higher-level themes. These higher-level themes described the larger concepts tying the lower-level codes together. For example, we discussed how "privacy," "surveillance," "unsafe content," and "data misuse" are tied together under the higher-level thematic concept of "safety." We defined this theme as a risk that data transparency might pose, particularly to data subjects and the public. The application of themes was discussed with the team and further refined.

Overall, we coded our documents using tiered codes. At the highest level were two sentiments: risk or benefit. Beneath these two sentiments, we coded eight higher-level themes: four risks and four benefits. Nested beneath these higher-level codes, we developed more specific lower-level codes associated with each risk and benefit: seven described risks and eight described benefits. Not all higher-level themes were broken down into lower-level codes. For example, the higher-level theme of "Contamination" did not have more specific lower-level codes; meanwhile, the higher-level theme of "Accountability" had three lower-level codes: "advocacy," "ownership," and "oversight." Our codebook can be found at https://doi.org/10.5281/zenodo.18340221.

The final output of our analysis is a taxonomy of four risks and four benefits associated with AI data transparency. We describe our step-by-step coding process in Table 1. Though we present descriptive proportions of the benefits and risks we identified in Section 4, our overall approach is Interpretivist in nature [152]. Given the Interpretivist nature of this work, many of the concepts we identify in our taxonomy are not isolated; much like all of social reality, they have a tendency to overlap or reflect multiple worldviews [52]. As such, we abide by the tenets of Interpretivist qualitative research and did not seek to quantitatively define agreement [144, 195]. Agreement on the themes presented in this work was reached through team discussions, leading to iterative refinement (see Table 1). We still choose to report descriptive statistics in our findings, so that readers can get a sense of how often sources expressed risks and benefits. In particular, we report how often we coded sources as expressing risks or expressing benefits. We remind readers that these statistics are still rooted in authors' interpretations about what constituted a risk or a benefit.

## 4 A Taxonomy of Risks and Benefits of AI Data Transparency

We now present our taxonomy of risks and benefits of AI data transparency, together with illustrative examples from our analysis. We first present four risks of data transparency: scrutiny, contamination, competitiveness, and safety. We then present four benefits of data transparency: accountability, innovation, integrity, and suitability. We also present two factors central to determining the implications of AI data transparency: stakeholder positions and transparency modalities. Definitions for risks and benefits can be found in 2. Definitions for each factor can be found in 3.

### 4.1 Risks of AI Data Transparency

Risks made up 40.4% of our coding, with 52.9% of sources expressing one or more risks of data transparency. Many risks were articulated by the developers of FMs and their associated dataset. 36% of developer sources expressed one or more risks, making up 58% of all developer codes. Meanwhile, 62.6% of researchers expressed one or more risks[5]; 55.6% of law and policy leaders expressed one or more risks; and 33.3% of NGOs expressed one or more risks. Many researchers and law and policy leaders also seemed aware of the arguments made by some developers against data transparency, such as how data transparency could negatively affect the competitiveness of developers and more technical researchers. Many of the risks were associated with fully or partially releasing data, but also

[5]We note that those sources sampled from Google Scholar made up the majority (37%) of risk codes, with 34% of Scholar sources articulating one or more risks. While sampled to represent researchers broadly, this may also indicate a diversity of researchers, who may be focused on development or protecting technical breakthroughs.

| *Step* | *Description* | *Result* |
|---|---|---|
| Familarization | Coded a random sample of excerpts with descriptions of how transparency was being described in the text. Discussed themes we were seeing relevant to transparency risks and harms. | A better shared understanding of how transparency is described in the documents. The development of a guiding question for more focused coding: What does this statement say or instinuate that transparency enables? |
| Open Coding | Coded each excerpt with short representative labels of a benefit and/or risk (e.g., oversight, reproducability). | 21 codes divided into 10 risks (liability, backlash, cost, competitiveness, feasibility, privacy surveillance, safety misuse, contamination) and 11 benefits (evaluation, reproducibility, collaboration, participation, innovation, advocacy, ownership, oversight, suitability, trustworthiness). |
| Discussion of Initial Codes | Discussed existing open codes. Developed thematic grouping of open codes by discussing their similarities. Grouped open codes into higher-level codes that were descriptive of the 21 lower-level codes. | 8 themes divided into 4 risks (scrutiny, contamination, competitiveness, safety) and 4 benefits (accountability, innovation, integrity, suitability). |
| Defining Themes | Wrote a definition for each higher-level theme. Discussed and refined these definitions as a group. | Codebook with definitions. |
| Thematic Coding | Grouped excerpts under these 8 themes. | Final codebook with examples grouped into themes. |

**Table 1: A table describing our analysis process [170] step-by-step.**

| | *Theme* | *Risk / Benefit* | *Core Cause(s)* | *Implication(s)* |
|---|---|---|---|---|
| **Risks** | Contamination | Training data is compromised or tainted. | Fully open data availability. | Can result in poor performance, bias, or homogenization. |
| | Competitiveness | Failure or success in the AI marketplace. | Fully open data availability &/or Robust documentation. | Can cause a developer or company to lose their competitive edge in the AI marketplace. |
| | Safety | Failure to mitigate harm or injury due to data use. | Fully open data availability | Access to the data may cause harm or injury (e.g., through viewing unsafe content, privacy exposure, or the data being used maliciously) |
| | Scrutiny | Critical observation of data practices. | Fully open data availability &/or Robust documentation. | Developers or companies facing legal consequences or brand damages. |
| **Benefits** | Accountability | Ability to hold a company or developer accountable for data practices. | Partially or fully open data availability &/or Robust data documentation. | Advocacy on behalf of those implicated or harmed by data practices &/or Improved legal oversight and policy and lawmaking activities given increased understanding of data practices &/or Insights into data provenance that improve recourse and control over data for data owners and subjects |
| | Innovation | Improve SOTA and introduce new AI methods, products, use cases, etc. | Fully open data availability. | Access to data will help smaller, less resourced, more diverse individuals and organizations participate in the AI landscape and thus improve the AI marketplace. |
| | Integrity | Ability to assess ethical principles and claims underlying both datasets and models. | Partially or fully open data availability &/or Robust documentation. | Access to specific parts of the data or robust documentation about specific aspects of the data can allow stakeholders to evaluate the ethical principles of datasets and models & Access to data can allow stakeholders to reproduce scientific results. |
| | Suitability | Ability to assess whether a model is suitable for intended use. | Robust documentation. | Users can use information about the dataset to assess whether a model is suitable for their use cases. |

**Table 2: A table describing each of the risks and benefits of FM data transparency, the core cause(s) of the risk/benefit (i.e., data documentation vs. data availability), and what specific implications the risk/benefit may enable.**

with documentation being too comprehensive or detailed. It is also notable that many risks overlap, meaning that "contamination" can impact "competitiveness," for example. We show example overlaps in Figure 3.

*4.1.1 Contamination.* Approximately 5.2% of sources we analyzed contained concerns about *sharing data* due to potential **data contamination**, making it the least expressed risk in our corpus. Data

contamination occurs when the test dataset is inadvertently included in the training dataset. This risk is primarily significant when sharing carefully curated evaluation datasets, especially those not available online, as they could be absorbed into the training data for FMs during (re)training or fine-tuning. As FMs are frequently updated with new data, the likelihood of contamination grows. Evaluation datasets intended to assess a model's generalization ability may instead **influence its performance**, undermining reliability and inflating results. As the developers of GPT-3 wrote: *"A major methodological concern with language models pretrained on a broad swath of internet data, particularly large models with the capacity to memorize vast amounts of content, is potential contamination of downstream tasks by having their test or development sets inadvertently seen during pre-training."* [54] They state that even a small bug in filtering can lead to contamination that is too costly to fix via retraining. The developers of Gemini 1.5 similarly wrote, *"We found controlling for accidental leakage on webpages and open-source code repositories to be a non-trivial task, even with conservative filtering heuristics."* [204]. Filtering out overlapping data can become increasingly complex if the same datasets are regularly released and used by model developers.

Contamination is primarily an issue when the dataset itself is released publicly. However, issues of **homogenization** can occur even if training data is not released publicly, as developers source data from the same publicly available sources and datasets. As Hopkins et al. write, homogenization *"can result in the long tail of more creative, infrequent, or context-specific expressions effectively disappearing over time"* [103]. The release of more proprietary datasets may potentially increase homogenization if other developers start to regularly use the same released datasets for training and fine-tuning.

*4.1.2 Competitiveness.* About 23.5% of sources we analyzed contained statements about how comprehensive and transparent documentation or data sharing may risk developers' **competitive advantage** in the burgeoning AI marketplace. Competitiveness refers to how a developer and/or their company can achieve success in the face of direct market competition. Data is thus treated as a trade secret, given its centrality in FM development. As Alderman et al. report, *"In documentation published at the launch of its GPT-4 model, OpenAI (2023) stated that it would not share detailed information about "data set construction" and other aspects of the model's development due to 'the competitive landscape and the safety implications of large-scale models'"* [21]. Transparency advocates, including law and policy stakeholders, also acknowledge that *"transparency may compromise competitive advantage or intellectual property rights"* [47].

Though many developers who published their models to HuggingFace were asked questions in the community forums about their training data, we observed one instance where a developer openly responded. One of the developers of Mixtral 7B [109] responded to the user question "Could you please share some of [information on the data used to train the baseline model] as well, in line with the model's open source philosophy?" with: *"Unfortunately we're unable to share details about the training and the datasets (extracted from the open Web) due to the highly competitive nature of the field. We appreciate your understanding!"*

Interestingly, even stakeholders who otherwise support data sharing expressed some concerns about competitiveness. For example, Tseng et al. found in their co-design exploration of a participatory LLM for journalism that the journalists they interviewed still saw the need to protect proprietary data [213]. In such cases, the participants in Tseng et al.'s study felt that some data should be shared or pooled with other organizations, while more sensitive data should be kept secret (not solely for competitive purposes, but also for safety purposes).

Given the centrality of data to AI development, competitiveness is threatened not only by releasing data to the public, but by robust documentation that reports in great detail the collecting, annotating, and cleaning processes of data.

*4.1.3 Safety.* About 39.2% of sources we analyzed expressed concerns about **safety issues** associated with transparency. Safety broadly refers to mitigating the likelihood that data can cause harm or injury. Safety issues were depicted as impacting exploited data subjects, data users who may be exposed to unsafe data, and the public who may be targets of malicious data use. NIST wrote that FMs present *"eased production of and access to violent, inciting, radicalizing, or threatening content as well as recommendations to carry out self-harm or conduct illegal activities"* [154]. Developers thus try to eliminate unsafe data prior to training and fine-tuning so that models do not learn and regurgitate harmful responses (e.g., [89, 166, 204, 205]). As complete removal of harmful content is rarely guaranteed, safety risks remain if developers release a dataset that could still contain unsafe content or personally identifiable information (as seen with LAION [43, 164, 209] and ImageNet [44, 67]). Safety risks impact data users who are exposed to unsafe or offensive content.

Safety risks can also be consequential for unsuspecting data subjects whose privacy-sensitive content is exposed. For example, Franchi et al. found that datasets with dense street imagery regularly fail to protect identities even with face and license plate anonymization techniques. They recommended that the organizations sharing such privacy-sensitive datasets consider how best to *"take responsibility for ensuring their ethical use by researchers, governments, and corporations"* given that *"unconditional sharing, without legal repercussions, will inevitably cause privacy harm to groups"* [81]. In a single instance, one of the developers of Guanaco openly criticized someone reporting the dataset to HuggingFace for their exposure of private data. Rather than removing private data from the dataset, they removed the entire dataset, updating the data card with the statement, *"The people here don't deserve it"* [63]. Publicizing a dataset opens it up to external scrutiny that may also reveal unsafe data practices.

Finally, some authors expressed concerns that being too transparent or open about data could lead to malicious use. Releasing data, or even documenting some forms of data curation in extensive detail, may allow malicious actors to use that information or data in harmful ways [20, 123, 142, 161].

*4.1.4 Scrutiny.* About 17.6% of sources we analyzed contained broad concerns about **scrutiny** from external stakeholders enabled by dataset transparency. At the forefront of these concerns are that developers, usually in commercial contexts, may face **legal consequences**, especially amid regulatory ambiguity and ongoing

lawsuits related to IP protections [121], particularly related to harms for copyright holders in generative AI contexts[137, 139]:

> *"Existing foundation models are trained on copyrighted material. Deploying these models can pose both legal and ethical risks when data creators fail to receive appropriate attribution or compensation ... Thus, the risk of infringement is real, and fair use will not cover every scenario where a foundation model is created or used. The exact amount of risk is unclear, and the law will evolve with ongoing litigation"* [86].

Legal risks are also associated with the presence of undetected illegal content in massive datasets. While developers describe attempts to remove illegal and harmful content from their datasets [89, 204], it is possible that the removal of some content is missed. For the open-source developer community, this is particularly pertinent, as *"liability for harms arising from downstream usage could chill the open FM ecosystem by exposing open FM developers to severe liability risk"* [46].

Commercial developers are also concerned that opening up about their data practices could **damage their brand reputation** and thus cause them to incur losses. As Pushkarna et al. write in [165], *"Any information included in a transparency artifact can be expected to receive greater scrutiny."* Such reputational harm can be seen in instances where companies have released datasets intended for fairness purposes, but were scrutinized for other ethical deficiencies, like subject consent (e.g., [93, 148]).

Commercial developers, in particular, may perceive opacity about data curation as a form of protection from legal risks and brand damages: *"Model developers that transparently disclose and openly provide data are subject to greater risk than developers that obfuscate the data they use, even if the underlying facts are identical"* [46]. However, especially given the increase in laws requiring transparency (e.g., [9]), opacity may be insufficient to protect developers from legal risks, including if they have trained on privacy-violating data, data to which they have not secured appropriate rights, or otherwise problematic data.

## 4.2 Benefits of AI Data Transparency

Benefits made up 59.6% of our coding (as opposed to 40.4% dedicated to risks), with 96.9% of sources expressing one or more benefits of transparency (as opposed to 52.9% who expressed one or more risks). Benefits of AI data transparency were largely expressed in researcher, NGO, and law and policy sources. These stakeholders also saw data transparency as affecting the general public. 100% of NGOs expressed one or more benefits; 88.8% law and policy leaders expressed one or more benefits; and 81.3% of researchers expressed one or more benefits. Meanwhile, 42% of developers expressed one or more benefits. While these stakeholder groups promoted more robust, transparent documentation, some researchers, NGOs, and community developers additionally advocated for fully open-source datasets (e.g., [11, 133, 229]). It is also notable that many of these benefits overlap, meaning that "accountability" can impact "integrity," for example. We show example overlaps in Figure 3.

*4.2.1 Accountability.* About 35.9% of sources contained arguments that more transparency about data practices can ensure **accountability** for irresponsible, unethical, or illegal data practices. On the one hand, transparency might allow **advocacy** on behalf of those implicated by extractive data practices. For example, that transparency into data sources could allow law and policy leaders to craft stronger arguments for provenance and copyright laws [48, 86, 239]; transparency into the labor practices underlying dataset curation could allow both affected workers and regulators to advocate for better working conditions [1, 47, 128]; and transparency into the environmental factors caused by collecting data and training models could lead to more robust environmental policies around AI [76, 239].

Besides advocating for better or new policies, transparency has also been seen as a way to enable regulators to hold developers accountable via **oversight**. For example, the explanatory notice for the general-purpose AI training data template by the EU Commission shows how transparency is intended to facilitate compliance with the EU AI Act and broader rights:

> *"Transparency of the training data in the Summary may facilitate data subjects' rights and more broadly support the enforcement of the Union data protection rules"* [14].

Transparency was also seen as enabling ethical oversight over concerns with data **ownership** (e.g., [13, 14, 92]). Currently, a lack of transparency is seen as a barrier to knowing where data comes from and who produced that data, making it impossible for either rightholders or data subjects to oversee its use or hold developers accountable for its use. As the authors of the open-source model and dataset BLOOM state: *"Abstractive approach[es] to data curation leads to corpora that are difficult to meaningfully document and govern after the fact, as the provenance and authorship of individual items is usually lost in the process"* [234]. Transparency about data provenance should allow "*rightsholders [to] choose to reserve their rights over their works"* [14].

*4.2.2 Innovation.* About 34% of sources expressed access to data as central to **innovation**: improving existing AI techniques or developing new ones. Transparency, particularly in the form of open-source datasets, was of interest to *"less-well-resourced actors such as academic labs and open-source developers"* seeking to *"advance the AI capability frontier"* [184]. As the Open Source Initiative states, *"Open Source AI ... spurs innovation and quality due to increased competition and tackles AI monoculture by providing more stakeholders access to foundational technology"* [11].

At the core of promoting innovation was the drive to ensure that diverse stakeholders can **participate** in the field. Developers of the cross-institutional fully-open models and datasets, RedPajama, stated: *"In many ways, AI is having its Linux moment. Stable Diffusion showed that open-source can not only rival the quality of commercial offerings like DALL-E but can also lead to incredible creativity from broad participation by communities around the world"* [6]. Extending beyond many other open datasets that prohibit commercial use, the developers of RedPajama make their dataset available for commercial use specifically to promote innovation and "democratiz[e] access" [226].

*4.2.3 Integrity.* About 47.1% of sources promoted transparency so that external stakeholders can ensure the **integrity** of the data used to train models. This makes integrity the most ubiquitous benefit of

transparency in our findings. As George et al. write, *"comprehensive audits of training data [... are] a privilege currently reserved only for developers"* [86]. External stakeholders should be able to **evaluate** that the datasets underlying AI meets principles of fairness (i.e., fair distributions, a lack of bias towards certain properties or groups [11, 174, 215]), privacy (i.e., data subjects are properly anonymized and cannot be easily de-anonymized [58, 203]), and safety (i.e., there is no harmful content, like violence or child exploitation imagery (CEI or CSAM [57, 164, 203–205]), or offensive content [166, 203, 246]):

> *"Understanding the origins and characteristics of the training data is crucial for assessing the reliability and biases inherent in LLMs. A lack of transparency about data sources and composition hinders the ability to identify and mitigate biases which can be perpetuated in model outputs"* [133].

Researchers, in particular, regularly expressed concern that data opacity is a primary barrier to **reproducibility** (e.g., [19, 106, 238]), a major principle underpinning the integrity of scientific progress. Simson et al. lament the common practice of solely naming a dataset, without either sufficiently describing its composition or provenance, or providing a dataset for others to replicate scientific results:

> *"This is a significant risk to the reproducibility and generalization of ... research for a combination of two reasons: (1) many publications do not document their usage of a dataset sufficiently, assuming that merely the name of a dataset clearly identifies its usage and (2) publications that do document data usage or offer reproducible code vary greatly in their usage, disproving the idea that merely identifying a dataset by its name is sufficient information"* [194].

As a result of concerns over issues of integrity, some developers have chosen to release datasets. For example, the developers of OLMo released the dataset used to train it, called Dolma, alongside detailed data documentation. They stated that *"a core tenet of our work is openness, which we define to mean (i) sharing the data itself and (ii) documenting the process to curate it"* to *"enable the broader research community to use our artifacts to study (and scrutinize) language models being developed today, even those developed behind closed doors"* [196].

*4.2.4 Suitability.* About 9.2% of sources we examined described wanting more transparent information about datasets to know whether a model is **suitable** for specific uses. For example, if a developer wishes to build on a model with open weights, such as Llama, knowing information about the dataset can help them *"assess its suitability for their purpose and avoid misuse"* ([130]). Further, even if datasets are released for others to use, determining their utility for modeling still requires appropriate documentation to be provided. As Alderman et al. write:

> *"We believe that transparent communication of dataset composition, including biases and limitations, can mean AI ... technologies are developed with the most appropriate (rather than the most popular) datasets"* [21].

## 4.3 Factors Influential to Risk/Benefit Perceptions

In the prior section, we articulated how AI data transparency can entail certain risks *and* benefits. However, we observed that both risks and benefits are contingent on two factors that implicitly arose through our analysis: (1) the *positions* stakeholders took on AI data transparency; and (2) the *modality* of AI data transparency.

*4.3.1 Stakeholder Position: A Spectrum Between Transparency Advocacy vs. Transparency Opposition.* We observed stakeholders falling on a spectrum between two broad categories: (1) transparency advocacy, pushing for better AI data transparency, and (2) transparency opposition, resisting (often more implicitly or silently) AI data transparency. The two poles on this spectrum between advocacy and opposition were not inherently associated with specific types of stakeholders, though they did often correlate with specific stakeholder types. For example, we found that a more risk-aware approach to transparency was expressed by commercially-oriented developers and law and policy leaders. By contrast, those advocating for data transparency seemed less concerned by these risks. Researchers, NGOs, law and policy leaders, and more community-oriented developers largely associated transparency with benefits.

These positions do not exist on a binary of "anti" versus "pro" transparency stances. Some stakeholders may oppose transparency in some cases, while advocating for transparency in others. For example, law and policy leaders may push for transparency about the sourcing of data to better uphold legal frameworks, while recognizing that making certain data publicly available may create privacy concerns (e.g., the EU AI Act [14]). Therefore, it is important to consider that stakeholder positions on transparency can shift along a spectrum between these two poles, depending on both the modality transparency takes on and their interpretation of risks versus benefits. We acknowledge the spectrum between advocacy and opposition, but use the broad terms *transparency advocate* and *transparency opponent* for simplicity throughout the remainder of the paper.

Transparency advocates often centered notions of power between developers and the public or smaller organizations. Commercial developers, who curate and store datasets, are seen as having disproportionate power to make data transparency decisions—particularly given a lack of legal measures to impose data transparency. Developers can choose to share aspects of datasets that further their underlying ambitions while obfuscating aspects that might block those ambitions [244]. On the other hand, those who are not curating AI datasets themselves—such as independent researchers, NGOs, law and policy stakeholders, and developers in smaller organizations (or outside of organizations)—are positioned without as much power to make any transparency decisions. The power to reveal or obfuscate is seen as largely in the hands of commercial developers who have the capital to build large-scale training datasets, leaving even community developers interested in transparency benefits like innovation and suitability equally as deprived of such benefits as non-developers. The skew in transparency decision-making power is exemplified in the increasing tendency for corporate developers to *"release by blog post"* and thus bypass the critical scrutiny that scientific peer review entails [131].

| *Factor* | *Type* | *Definition* | *Spectrum* |
|---|---|---|---|
| **Stakeholder Position** | Transparency Advocate | A stakeholder who advocates for (some or all) AI data transparency due to the perceived benefits to their values, often dependent on transparency modality | May advocate for partial or complete data transparency (in the form of either documentation or availability) to enact benefits |
| | Transparency Opponent | A stakeholder who opposes (some or all) AI data transparency due to the perceived risks to their values, often dependent on transparency modality | May oppose either partial or complete data transparency (in the form of either documentation or availability) to reduce risks |
| **Transparency Modality** | Data Documentation | The descriptions of different elements relevant to an AI dataset | May include descriptions of data sources, data composition, data creation processes, etc. in differing levels of detail |
| | Data Availability | The availability of the (pre-training, fine-tuning, evaluation) data to external parties | May range from entirely closed-sourced (not available to external parties) to entirely open-source (completely available to external parties) |

**Table 3: A table describing the two factors shaping whether AI data transparency is a risk or a benefit: (1) stakeholder positions on data transparency and (2) the modality of data transparency. Stakeholder positions are not fixed nor necessarily binary and may occupy a spectrum between advocacy and opposition, depending on modality and perceived risks and benefits.**

*4.3.2 Transparency Modalities: Data Availability and Data Documentation.* The extent of these risks and benefits is further shaped by two transparency modalities: (1) data documentation and (2) data availability [229]. Data documentation refers to descriptions of different elements of a dataset, including its composition and creation process. Data availability refers to the accessibility of data for training, fine-tuning, and/or evaluation.

Much like with stakeholder positions, these two modalities of data transparency are not binary, but rather exist on a spectrum. **Data documentation is a spectrum between no documentation and robust documentation** on any contextually relevant dataset attributes. **Data availability is a spectrum between closed-source and open-source**, where fully closed-source means no access to any of the data and fully open-source means full access to all of the data, including relevant annotations or metadata. Further, a developer team may use multiple datasets for pre-training, fine-tuning, or evaluating their model, but choose only to document or release one or some (e.g., the developers of Nemotron-4 released their HelpSteer2-Preference dataset for model alignment [162], but not their pre-training data).

Both dataset documentation and dataset availability are also independent variables. It is possible to release an open-source dataset with no or poor documentation, just as it is possible to release an open-source dataset with robust documentation. While most commercial datasets are closed-source with poor documentation (e.g., GPT, Llama, Gemini), some developers have released open-source datasets with robust documentation (e.g., BLOOM, Aya).

Modality can influence whether a stakeholder is an advocate or opponent of AI data transparency. For example, developers may face increased scrutiny if they provide robust documentation about specific characteristics of the data, such as the specific sources of the data, regardless of whether they do or do not release the dataset. Robust documentation can elevate legal or reputational risks by inviting further scrutiny upon dataset characteristics, such as data collection processes, data worker treatment and wages, and the question of whether subjects have consented to be in the dataset. Meanwhile, releasing a fully open-source dataset containing sensitive or personally identifiable information, even without documentation, may expose developers to scrutiny about safety—even when other ethical aspects of the dataset are not disclosed, such as worker wages.

Given the enduring opacity around dataset documentation, most transparency proponents advocate for better documentation norms and practices. Transparency advocates have provided many tools for improving the robustness of dataset documentation, like guides about how and what to document (e.g., [84, 146, 175, 229]). Some have also argued that robust documentation, including dataset recipes (i.e., instructions for creating similar datasets), are sufficient for enabling benefits like integrity. For example, the NeurIPS Paper Checklist Guidelines for best practices in machine learning research states: *"Reproducibility can be accomplished in various ways ... In general. [sic] releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results ... or other means that are appropriate to the research performed"* [17].

Even while some transparency advocates, like NGOs, have called for open-source datasets, it is notable that not all transparency advocates ask for open-source datasets. We did not observe any calls for open-source datasets from law and policy leaders. Transparency advocates often advocated for open-source datasets when they were particularly concerned with achieving innovation or ensuring integrity via reproducibility. Others, less interested in development or scientific integrity, advocated instead for the ability of third parties to evaluate integrity in the form of safety or fairness checks, which may require time-restricted access to the data or access to a limited sample. For example, McKinney et al. stated: *"If a dataset cannot be shared with the entire scientific community, because of licensing or other insurmountable issues, at a minimum a mechanism should be set so that some highly trained, independent investigators can access the data and verify the analyses"* [145].

# 5 Discussion: Treading the Transparency Tightrope

While templates on *what* should be documented about AI data has been a crucial first step in data transparency efforts (e.g., with data cards [165] or datasheets [84]), it is also crucial that transparency advocates and opponents alike are able to articulate *why* and *for what purposes* AI data transparency—or opacity—is necessary. In this work, we turned to cross-disciplinary discussions about AI data transparency in order to conceptualize transparency as contingent [66]. We developed a taxonomy to codify the specific risks and benefits underlying current AI data transparency discourse. Whether data transparency is perceived as a risk of benefit depends on whether a stakeholder takes on the position of either data transparency advocate or data transparency opponent, though the degree to which a stakeholder may advocate or oppose can vary. Perceptions of risks and benefits are shaped by the specific modality data transparency takes on, whether in the form of data availability or data documentation.

We now show how transparency advocates can use our taxonomy to ground data transparency advocacy in risks and benefits. More specifically, we use our taxonomy to argue for *situational data transparency*—data transparency that is contingent upon how transparency presents benefits and risks to specific stakeholders with specific goals using specific modalities.

## 5.1 Situational Data Transparency: Using the Taxonomy to Maximize Situational Benefits and Minimize Situational Risks

In Section 4.3, we broke down how two factors influenced whether transparency presented a risk or a benefit: (1) the position of a stakeholder in relation to data transparency and (2) the modality that data transparency takes on. In this section, we seek to demonstrate through a theoretical example how an open-source dataset with robust documentation may present *situational benefits* to transparency advocates, but also *situational risks* to transparency opponents (see Figure 4). Through this example, we highlight key opportunities for transparency advocates to consider how best to optimize for situational benefits while mitigating situational risks.

In particular, we delve into how **decisions around the degree (ranging from partial to complete) of transparency modality (data documentation vs. data availability) may activate different benefits and risks**. We describe how the more robust the data availability, the more that benefits like "Accountability" and "Suitability" and risks like "Scrutiny" and "Competitiveness" may increase. Meanwhile, the more robust data documentation, the more that benefits like "Innovation" and "Integrity" and risks like "Contamination," "Safety," "Scrutiny," and "Competitiveness" may increase. On the other hand, no data availability and no data documentation may have no risks, but also no benefits. We argue that to best utilize situational transparency, transparency advocates must make convincing arguments that the situational risks do not outweigh situational benefits. For an advocate to better position the benefits of transparency, they should be able to articulate how the degree of a transparency modality maps to specific, contextualized benefits that outweigh or avoid relevant risks.

Such a scenario reflects common tensions we observed between transparency advocates in research (e.g., [92, 184, 234]), NGO (e.g., OSI), and open-source development communities (e.g., Cohere, BigScience, and even commenters on some proprietary model cards on HuggingFace [109]), versus those developers and law and policy experts in private companies (e.g., OpenAI, Anthropic, Meta). But we also note once more that the classifications of "advocate" and "opponent" are not binary or tied to specific types of stakeholders—some stakeholders may oppose certain forms of transparency to a certain degree, for example. However, we once more discuss them as simply "advocates" and "opponents" in this section to more clearly illustrate differential benefits and risks.

Our analysis throughout this section is represented in the visualization in Figure 4, which we will point to throughout these scenarios. By demonstrating how situational data transparency differs from transparency calls divorced from risks and from transparency obfuscation divorced from benefits, we seek to showcase how situational data transparency can offer a more pragmatic approach contingent upon stakeholder positions.

*5.1.1 Maximal Benefits to Transparency Advocates.* Drawing from the arguments advocating for data transparency in the sources we analyzed, we can imagine how maximal data availability and maximal data documentation may yield *maximal benefits*. For example, having complete access to the data underlying large proprietary LLMs, like GPT-4, could enable smaller-scale and open-source developers increased opportunities to *innovate*, given their own relative lack of resourcing [96, 179, 218]. Accurate and robust data documentation outlining data collection methods and sources, environmental impacts, bias assessments, and underlying labor could allow law and policymakers to hold the companies using the documented data *accountable* to relevant laws and best practices. Similarly, both full access to data and robust documentation could allow external researchers to ensure the *integrity* of released models by auditing datasets for principles of fairness and assessing the reproducibility of model results. Such access and documentation could also enable both law and policymakers and researchers to assess the *suitability* of the given data to the tasks a given model was trained to do, allowing them to make more informed decisions about what misuse of the data might constitute.

*5.1.2 Maximal Risks to Transparency Opponents.* Like with maximal benefits, we can also draw on the concerns sources expressed about data transparency to imagine how maximal data availability and maximal data documentation may yield *maximal risks*. For example, making data fully available is generally perceived as a risk to *competitiveness*; even documenting collection procedures to an extremely detailed degree can be perceived as a risk to competitiveness, as other companies may replicate the same methodologies, which has led developers to provide minimal documentation. Releasing the data also risks future *contamination* as model developers continuously aggregate data for model improvements. Given that so much data is collected from the web, some dataset developers may also oppose releasing the data due to potential *safety* issues with the content in the dataset. Further, it may introduce privacy concerns for any subjects whose personally identifiable data is in the dataset. Finally, making data fully available or robustly documenting data sourcing brings increased *scrutiny* to the data developers.

Due to their use of data infringing upon intellectual property rights [55, 110, 223], it is possible that scrutinizing these data and sources can carry a risk of financial loss if taken to court (e.g., [61, 200]).

Of course, not all risks to transparency opponents can be argued as justifiable. If a developer is concerned about scrutiny into data practices leading to legal consequences because they were fully aware that they engaged in illegal or questionable activities, then the risk they perceive about those practices is ethically (and potentially legally) fraught.

*5.1.3 Maximizing Benefits, Mitigating Risks.* Given that different stakeholders see risks and benefits to different transparency modalities, rather than focusing solely on maximizing transparency, transparency advocates should prioritize *situational data transparency* grounded in assessments of both benefits and risks. A more situational approach to data transparency was observed in certain sources we analyzed, which acknowledges both the risks and benefits of transparency. We already observed law and policy approaches that could be categorized as sitting in the "compromise zone" visualized in Figure 4. For example, we observed suggestions for mitigating safety issues for data subjects through privacy preservation and maintaining competitiveness for developers, while also enabling accountability measures, evaluation procedures, and innovation in the AI marketplace [14].

To best maximize benefits and mitigate risks, it is necessary to consider which modalities of transparency require data availability and documentation, what the degree of transparency for these modalities should be, and who those transparency modalities should be available to. For example, to best enable accountability to legal frameworks and ensure data is fair, it may be best to make data available to regulatory auditors, rather than the general public. At the same time, if the goal of transparency is to ensure data subjects can best take ownership of how their data is being used by developers, the general public may need robust documentation about the data sources used in a dataset or methods for assessing whether their data is present within a dataset—ideally, in a way that still enables safety. To ensure safety, a dataset developer may enable methods for subjects to search their presence in a dataset (e.g., by enabling search of specific usernames relevant to the sources used [97]) while foregoing releasing the data itself; they might consider only documenting methods to remove illegal, upsetting, or objectionable content, without making access to that content possible. Perhaps the developer may release only a portion of the data, which would enable benefits like innovation, integrity, accountability, and even suitability, but also maintain competitiveness and reduce contamination and safety concerns.

Overall, how best to make compromises between risks and benefits and promote situational data transparency depends on the stakeholders involved and how those stakeholders assess the risks and benefits of different transparency modalities. Using the taxonomy of risks and benefits, as well as the associated factors influential to them, different stakeholders can begin to ground their arguments in more concrete goals, rather than in vague calls that presume sweeping transparency as an inherent and necessary good [23, 66, 244].

## 6 Considerations for Promoting Situational Transparency

As we have shown, there are many considerations when weighing the risks and benefits of AI data transparency. Yet, many transparency advocates fail to explicitly name the benefits they seek to enact with transparency, leaving their intentions or desired actions implicit. Further, calls for transparency often fall short of considering potential risks, which can aid transparency advocates in pressuring transparency opponents (who might center, for example, competitiveness above all else).

Overall, we argue that there may be opportunities to promote benefits and bypass risks by strategizing about stakeholders and modalities. Our taxonomy presents a practical opportunity for transparency advocates (including HCI researchers) to understand tensions and consider trade-offs in their aim to increase AI data transparency. Beyond answering *why* datasets should be made more transparent, this taxonomy also presents a starting point for asking *how* to best make datasets more transparent—and *who* that transparency might benefit or harm. Building on these considerations, transparency advocates can take a step forward, following the position of Eyert and Lopez, to view transparency as a "communicative constellation" that centers the democratic negotiation between relevant stakeholders rather than technocentric unidirectional explainability mechanisms [78]. Beyond implementing data transparency, researchers can employ this taxonomy to more concretely define transparency, in both research approaches (e.g., in participatory research) and in situating findings and takeaways. As a starting point, we present six considerations for situational data transparency:

- **Consider your own position in relation to data transparency**: Identify the type of stakeholder you identify as and your reasons for advocating for data transparency. What benefits would data transparency enable for you or other stakeholders like you? What modalities would that transparency require to enable those benefits? Are there certain thresholds of transparency that would best enable those benefits? Understanding one's own position in relation to the perceived benefits of data transparency will aid in concretizing arguments for those benefits.
- **Identify other types of stakeholders**: Identify all stakeholders who will be impacted by data transparency decisions. Assess the benefits and the risks to each stakeholder. Consider whether benefits and risks will increase or decrease given documentation and availability decisions. Consider whether tensions between benefits and risks can be resolved by specifying different data modality regimes. For example, a dataset may be accessible to external auditing bodies or government bodies for safety evaluations, while being inaccessible to the general public [71]. Determining which stakeholders may or may not need certain types of data transparency can help make more grounded arguments about data transparency risks and benefits.
- **Provide incentives for the developers of datasets to be transparent**: Currently, much of the power to make transparency decisions is in the hands of developers. Outside of possible legal requirements (e.g., [9]), it may be necessary

to consider how best to incentivize data transparency, especially when proprietary data is considered a highly valuable asset. For example, numerous developers have embraced open-weight FMs because, for them, the perceived benefits outweigh possible risks. How can transparency advocates showcase the same benefits for data? Articulating the benefits of data transparency may require more grounded arguments that appeal to developer motivations. It is possible that some developers perceive the scrutiny of academic researchers as overly critical or untethered from production contexts, causing a freezing effect [182]. For example, HCI researchers might showcase how data transparency would benefit stakeholder groups that developers care about [192].

- **Show that transparency goals are necessary or successful**: Currently, transparency is largely measured in the form of absence/presence. For example, the Foundation Model Transparency Index [47] measures data transparency on whether or not decisions are documented. While in some cases, simply describing a decision-making process (e.g., when annotating data) may be considered sufficiently transparent, in other cases, access to the data itself may be necessary (e.g., for model innovations, for third-party performance evaluations). Transparency advocates must consider how to better measure or showcase that transparency about specific decisions can enable desired benefits. For example, user-centered design and evaluation methods can be used to assess the viability of transparency approaches (e.g., [75]. Similarly, advocates should consider when data opacity decisions may better align with broader responsible goals like fairness (e.g., deciding not to release a subset of the dataset to protect at-risk data subjects). For HCI researchers, like those in the CHI community, our taxonomy can be used to ground human-centered research with data subjects in exploring and advocating for more actionable transparency interventions that align with specific community goals (e.g., [68]).
- **Provide a specific list of transparency requests and why they matter:** Many calls for data transparency are vague in nature, arguing that transparency is necessary without identifying the modality of transparency that is desired or for what purposes it is necessary. Transparency advocates can better achieve their goals by specifying what types of transparency they desire. For example, the European commission recently published a template for developers of "general-purpose AI" that provides specific, purpose-driven requests for data documentation to enable accountability, integrity, and innovation goals while protecting competitiveness, such as: (1) general information (e.g., details about provider, model, types of training content, general characteristics); (2) a list of data sources; and (3) relevant data processing aspects (including details about the removal of illegal content) [14]. Researchers could similarly ground transparency advocacy in specific, actionable requests for change, as seen in calls for improving fairness measurement techniques (e.g., [108]) and categorical definitions (e.g., [222].
- **Push for explicit risk-based justifications for a lack of transparency**: Currently, most risk-based arguments for non-transparency lack true justification. Many developers of FM datasets did not actually discuss, even implicitly, their reasoning for not being more transparent, even in documenting the processes or composition underlying their datasets. If transparency about data is risky, then a developer should show how and why it is risky. Further still, a developer should be able to explain which aspects of the data present specific risks, as well as whether risks are tied to documentation or availability. For example, if a developer does not want to release a dataset because they claim it will expose data subjects to privacy concerns, why have they not enacted privacy-preserving techniques before releasing the data? Transparency advocates can rely on the taxonomy to interrogate whether transparency opponents are actually making reasonable arguments for their lack of transparency.

## 7 Limitations and Future Work

While we identified four risks and four benefits of FM dataset transparency, these are unlikely to be exhaustive. Our aim is to provide an initial taxonomy of the risks and benefits of data transparency, which are still absent from transdisciplinary responsible AI literature. Future research, particularly empirical studies involving diverse stakeholder groups (e.g., amongst data subjects), could uncover additional risks and benefits—potentially expanding on the stakeholder categories proposed by Eyert and Lopez [78]. We see many opportunities for future work that centers underrepresented stakeholders, like data subjects, NGOs, and community developers, using methodologies like participatory design, interviews, or social media data analysis. Further, the risks and benefits captured in this work are broad. As such, the variance or differential importance of each risk or benefit is not captured. For example, legal rather than reputational risks may be of greater concern to certain stakeholders than other risks. Further work is needed to empirically verify how stakeholder perspectives on risks and benefits map to the factors of transparency we identified. We also hope that the factors can inspire future work focused on determining the appropriate modality of transparency given dataset context and which stakeholder types will be using them, developing better incentives for developers to be transparent, and measuring whether transparency successfully enables benefits and mitigates risks. Further still, as even well-documented or openly available datasets are modified and reused across the ever-shifting landscape of AI, sometimes being made available and sometimes being retracted, determining how the benefits and risks associated with specific datasets might transform over time is an open question [64]. We also note that there may be continual regulatory interest pertaining to data transparency, given the anticipated increase in data transparency-related compliance over the next year (e.g., [9]). In presenting this taxonomy, we are not making any recommendations about how best to comply with these laws.

Finally, we observed that the four benefits we identified were often discussed in the sources we analyzed as increasing the **trustworthiness** of datasets, FMs, and their developers [80, 91, 220]. However, trustworthiness is similar to transparency in that it can enable other benefits (e.g., perceiving a model as suitable to adopt or perceiving an academic contribution as scientifically sound) but not

itself be the primary benefit [168]. For example, we observed developers reporting aspects of ethical and legal compliance to increase trust in their *accountability* to law and ethical principles (e.g., IBM's proprietary dataset for training Granite [10]). We believe future work focused explicitly on understanding what trustworthiness means is still needed, attending to questions about what *types* of trust transparency can enable and *who* that trust may benefit or harm, depending on what attributes trust is placed in.

## 8 Conclusion

The opacity of AI datasets has led many to advocate for increased transparency. Yet transparency is often portrayed as a primary virtue, without specificity about what benefits it enables—or what risks should be accounted for. In this work, we developed a taxonomy of risks and benefits associated with AI data transparency. Our taxonomy is intended to aid transparency advocates in making more informed *situational data transparency* demands by considering both the risks *and* benefits data transparency may pose to different stakeholders, including transparency opponents. Currently, power in making transparency decisions is skewed towards those transparency opponents who develop and own proprietary models and associated proprietary datasets, some of whom have strategically wielded opacity to maximize their own goals. We argue that, by mapping out both risks and benefits, advocates can better develop action plans to shift the balance towards maximizing benefits. We further provision transparency advocates with two modalities of data transparency—data documentation and data availability—so that they can better articulate what *kind* of transparency best advances their goals.

## Acknowledgments

We'd like to thank Tiffany Georgievski for her feedback and advice.

## References

[1] [n. d.]. The Transparency Coalition. https://www.transparencycoalition.ai.
[2] 2018. Open Data Sharing by Governments Is Stalling. https://internethealthreport.org/2018/open-data-sharing-by-governments-is-stalling/.
[3] 2021. AI Risk Management Framework. *NIST* (July 2021).
[4] 2023. Hiroshima Process International Code of Conduct for Organizations Developing Advanced AI Systems (30/10/2023).
[5] 2023. Interim Measures for the Management of Generative Artificial Intelligence Services.
[6] 2023. RedPajama, a Project to Create Leading Open-Source Models, Starts by Reproducing LLaMA Training Dataset of over 1.2 Trillion Tokens.
[7] 2023. Supporting Open Source and Open Science in the EU AI Act.
[8] 2024. Frontier AI Safety Commitments, AI Seoul Summit 2024.
[9] 2024. Generative Artificial Intelligence: Training Data Transparency, AB-2013, California Legislature (2023–24).
[10] 2024. Granite Foundation Models.
[11] 2024. The Open Source AI Definition – 1.0. https://opensource.org/ai/open-source-ai-definition.
[12] 2024. A Pro-Innovation Approach to AI Regulation.
[13] 2024. Recommendation of the Council on Artificial Intelligence.
[14] 2024. REGULATION (EU) 2024/1689 OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL.
[15] 2025. Commission Presents Template for General-Purpose AI Model Providers to Summarise the Data Used to Train Their Model.
[16] 2025. Initial Rescissions Of Harmful Executive Orders And Actions.
[17] 2025. NeurIPS Paper Checklist Guidelines. https://neurips.cc/public/guides/PaperChecklist.
[18] 2025. Template for the Public Summary of Training Content for General-Purpose AI Models.
[19] Rediet Abebe, Kehinde Aruleba, Abeba Birhane, Sara Kingsley, George Obaido, Sekou L. Remy, and Swathi Sadagopan. 2021. Narratives and Counternarratives on Data Sharing in Africa. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT '21)*. Association for Computing Machinery, New York, NY, USA, 329–341. https://doi.org/10.1145/3442188.3445897
[20] Emmanouil Adamakis, Michael Boch, Alexandros Bampoulidis, George Margetis, Stefan Gindl, and Constantine Stephanidis. 2023. Visualizing the Risks of De-anonymization in High-Dimensional Data. In *Information Technology and Systems*, Álvaro Rocha, Carlos Ferrás, and Waldo Ibarra (Eds.). Springer International Publishing, Cham, 27–37. https://doi.org/10.1007/978-3-031-33258-6_3
[21] Joseph E. Alderman, Maria Charalambides, Gagandeep Sachdeva, Elinor Laws, Joanne Palmer, Elsa Lee, Vaishnavi Menon, Qasim Malik, Sonam Vadera, Melanie Calvert, Marzyeh Ghassemi, Melissa D. McCradden, Johan Ordish, Bilal Mateen, Charlotte Summers, Jacqui Gath, Rubeta N. Matin, Alastair K. Denniston, and Xiaoxuan Liu. 2024. Revealing Transparency Gaps in Publicly Available COVID-19 Datasets Used for Medical Artificial Intelligence Development—a Systematic Review. *The Lancet Digital Health* 6, 11 (Nov. 2024), e827–e847. https://doi.org/10.1016/S2589-7500(24)00146-8
[22] Saghir Alfasly, Peyman Nejat, Sobhan Hemati, Jibran Khan, Isaiah Lahr, Areej Alsaafin, Abubakr Shafique, Nneka Comfere, Dennis Murphree, Chady Meroueh, Saba Yasir, Aaron Mangold, Lisa Boardman, Vijay Shah, Joaquin J. Garcia, and H. R. Tizhoosh. 2023. When Is a Foundation Model a Foundation Model. https://doi.org/10.48550/arXiv.2309.11510 arXiv:2309.11510 [cs]
[23] Emmanuel Alloa and Dieter Thomä. 2018. *Transparency, Society and Subjectivity: Critical Perspectives*. Springer.
[24] Melany Amarikwa. 2024. Internet Openness at Risk: Generative AI's Impact on Data Scraping. https://doi.org/10.2139/ssrn.4723713 social science research network:4723713
[25] Mike Ananny and Kate Crawford. 2018. Seeing without Knowing: Limitations of the Transparency Ideal and Its Application to Algorithmic Accountability. *New Media & Society* 20, 3 (March 2018), 973–989. https://doi.org/10.1177/1461444816676645
[26] Jerone Andrews, Dora Zhao, William Thong, Apostolos Modas, Orestis Papakyriakopoulos, and Alice Xiang. 2023. Ethical Considerations for Responsible Data Curation. In *Thirty-Seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.
[27] Ariful Islam Anik and Andrea Bunt. 2021. Data-Centric Explanations: Explaining Training Data of Machine Learning Systems to Promote Transparency. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems (CHI '21)*. Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3411764.3445736
[28] Amanda Askell, Miles Brundage, and Gillian Hadfield. 2019. The Role of Cooperation in Responsible AI Development. https://doi.org/10.48550/arXiv.1907.04534 arXiv:1907.04534 [cs]
[29] Sumit Asthana, Jane Im, Zhe Chen, and Nikola Banovic. 2024. "I Know Even If You Don't Tell Me": Understanding Users' Privacy Preferences Regarding AI-based Inferences of Sensitive Information for Personalization. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems (CHI '24)*. Association for Computing Machinery, New York, NY, USA, 1–21. https://doi.org/10.1145/3613904.3642180
[30] Pierre Azoulay, Joshua L. Krieger, and Abhishek Nagaraj. 2024. Old Moats for New Models: Openness, Control, and Competition in Generative AI. https://doi.org/10.3386/w32474 national bureau of economic research:32474
[31] Stefan Baack. 2024. A Critical Analysis of the Largest Source for Generative AI Training Data: Common Crawl. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency (FAccT '24)*. Association for Computing Machinery, New York, NY, USA, 2199–2208. https://doi.org/10.1145/3630106.3659033
[32] Kenneth D. Bailey. 1994. *Typologies and Taxonomies: An Introduction to Classification Techniques*. SAGE.
[33] Alistair Barr. 2023. Llama Copyright Drama: Meta Stops Disclosing What Data It Uses to Train the Company's Giant AI Models. https://www.businessinsider.com/meta-llama-2-data-train-ai-models-2023-7.
[34] Teanna Barrett, Chinasa T. Okolo, B. Biira, Eman Sherif, Amy Zhang, and Leilani Battle. 2025. African Data Ethics: A Discursive Framework for Black Decolonial AI. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency (FAccT '25)*. Association for Computing Machinery, New York, NY, USA, 334–349. https://doi.org/10.1145/3715275.3732023
[35] Adrien Basdevant, Camille François, Victor Storchan, Kevin Bankston, Ayah Bdeir, Brian Behlendorf, Merouane Debbah, Sayash Kapoor, Yann LeCun, Mark Surman, Helen King-Turvey, Nathan Lambert, Stefano Maffulli, Nik Marda, Govind Shivkumar, and Justine Tunney. 2024. Towards a Framework for Openness in Foundation Models: Proceedings from the Columbia Convening on Openness in Artificial Intelligence. https://doi.org/10.48550/arXiv.2405.15802 arXiv:2405.15802 [cs]

[36] Dan Bateyko and Karen Levy. 2025. One Bad NOFO? AI Governance in Federal Grantmaking. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency (FAccT '25)*. Association for Computing Machinery, New York, NY, USA, 1640–1652. https://doi.org/10.1145/3715275.3732109

[37] Austin Beacham, Emilie M. Hafner-Burton, and Christina J. Schneider. 2024. The Weaponization of Information Technologies and Democratic Resilience. (Nov. 2024).

[38] Elisa Bertino. 2020. The Quest for Data Transparency. *IEEE Security & Privacy* 18, 3 (May 2020), 67–68. https://doi.org/10.1109/MSEC.2020.2980593

[39] Elisa Bertino, Shawn Merrill, Alina Nesen, and Christine Utz. 2019. Redefining Data Transparency: A Multidimensional Approach. *Computer* 52, 1 (Jan. 2019), 16–26. https://doi.org/10.1109/MC.2018.2890190

[40] Eshta Bhardwaj, Harshit Gujral, Siyi Wu, Ciara Zogheib, Tegan Maharaj, and Christoph Becker. 2024. Machine Learning Data Practices through a Data Curation Lens: An Evaluation Framework. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency (FAccT '24)*. Association for Computing Machinery, New York, NY, USA, 1055–1067. https://doi.org/10.1145/3630106.3658955

[41] Clare Birchall. 2011. Introduction to 'Secrecy and Transparency': The Politics of Opacity and Openness. *Theory, Culture & Society* 28, 7-8 (Dec. 2011), 7–25. https://doi.org/10.1177/0263276411427744

[42] Abeba Birhane, Sepehr Dehdashtian, Vinay Prabhu, and Vishnu Boddeti. 2024. The Dark Side of Dataset Scaling: Evaluating Racial Classification in Multimodal Models. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency (FAccT '24)*. Association for Computing Machinery, New York, NY, USA, 1229–1244. https://doi.org/10.1145/3630106.3658968

[43] Abeba Birhane, Vinay Prabhu, Sanghyun Han, Vishnu Boddeti, and Sasha Luccioni. 2023. Into the LAION's Den: Investigating Hate in Multimodal Datasets. *Advances in Neural Information Processing Systems* 36 (Dec. 2023), 21268–21284.

[44] Abeba Birhane and Vinay Uday Prabhu. 2021. Large Image Datasets: A Pyrrhic Win for Computer Vision?. In *2021 IEEE Winter Conference on Applications of Computer Vision (WACV)*. 1536–1546. https://doi.org/10.1109/WACV48630.2021.00158

[45] Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, Erik Brynjolfsson, S Buch, Dallas Card, Rodrigo Castellon, Niladri S Chatterji, Annie S Chen, Kathleen A Creel, Jared Davis, Dora Demszky, Chris Donahue, Moussa Doumbouya, Esin Durmus, Stefano Ermon, John Etchemendy, Kawin Ethayarajh, Li Fei-Fei, Chelsea Finn, Trevor Gale, Lauren E Gillespie, Karan Goel, Noah D Goodman, Shelby Grossman, Neel Guha, Tatsunori Hashimoto, Peter Henderson, John Hewitt, Daniel E Ho, Jenny Hong, Kyle Hsu, Jing Huang, Thomas F Icard, Saahil Jain, Dan Jurafsky, Pratyusha Kalluri, Siddharth Karamcheti, Geoff Keeling, Fereshte Khani, O Khattab, Pang Wei Koh, Mark S Krass, Ranjay Krishna, Rohith Kuditipudi, Ananya Kumar, Faisal Ladhak, Mina Lee, Tony Lee, Jure Leskovec, Isabelle Levent, Xiang Lisa Li, Xuechen Li, Tengyu Ma, Ali Malik, Christopher D Manning, Suvir P Mirchandani, Eric Mitchell, Zanele Munyikwa, Suraj Nair, Avanika Narayan, Deepak Narayanan, Benjamin Newman, Allen Nie, Juan Carlos Niebles, Hamed Nilforoshan, J F Nyarko, Giray Ogut, Laurel Orr, Isabel Papadimitriou, Joon Sung Park, Chris Piech, Eva Portelance, Christopher Potts, Aditi Raghunathan, Robert Reich, Hongyu Ren, Frieda Rong, Yusuf H Roohani, Camilo Ruiz, Jack Ryan, Christopher R'e, Dorsa Sadigh, Shiori Sagawa, Keshav Santhanam, Andy Shih, Krishna Parasuram Srinivasan, Alex Tamkin, Rohan Taori, Armin W Thomas, Florian Tramèr, Rose E Wang, William Wang, Bohan Wu, Jiajun Wu, Yuhuai Wu, Sang Michael Xie, Michihiro Yasunaga, Jiaxuan You, Matei A Zaharia, Michael Zhang, Tianyi Zhang, Xikun Zhang, Yuhui Zhang, Lucia Zheng, Kaitlyn Zhou, and Percy Liang. 2021. On the Opportunities and Risks of Foundation Models. *ArXiv* abs/2108.0 (2021).

[46] Rishi Bommasani, Sayash Kapoor, Kevin Klyman, Shayne Longpre, Ashwin Ramaswami, Daniel Zhang, Schaake Marietje, Daniel E. Ho, Arvind Narayanan, and Percy Liang. 2023. *Considerations for Governing Open Foundation Models*. Technical Report. Stanford University's Institute on Human-Centered Artificial Intelligence (HAI).

[47] Rishi Bommasani, Kevin Klyman, Shayne Longpre, Sayash Kapoor, Nestor Maslej, Betty Xiong, Daniel Zhang, and Percy Liang. 2023. The Foundation Model Transparency Index. https://doi.org/10.48550/arXiv.2310.12941 arXiv:2310.12941 [cs]

[48] Rishi Bommasani, Kevin Klyman, Shayne Longpre, Betty Xiong, Sayash Kapoor, Nestor Maslej, Arvind Narayanan, and Percy Liang. 2024. Foundation Model Transparency Reports. https://doi.org/10.48550/arXiv.2402.16268 arXiv:2402.16268 [cs]

[49] Rishi Bommasani, Daniel Zhang, Tony Lee, and Percy Liang. 2023. *Improving Transparency in AI Language Models: A Holistic Evaluation.* Technical Report. Stanford HAI.

[50] Giovanna Borradori. 2016. Between Transparency and Surveillance: Politics of the Secret. *Philosophy & Social Criticism* 42, 4-5 (May 2016), 456–464. https://doi.org/10.1177/0191453715623321

[51] Mohamed Boukhlif, Nassim Kharmoum, and Mohamed Hanine. 2024. LLMs for Intelligent Software Testing: A Comparative Study. In *Proceedings of the 7th International Conference on Networking, Intelligent Systems and Security (NISS '24)*. Association for Computing Machinery, New York, NY, USA, 1–8. https://doi.org/10.1145/3659677.3659749

[52] Geoffrey C. Bowker and Susan Leigh Star. 1999. *Sorting Things Out: Classification and Its Consequences.* MIT Press.

[53] Blake Brittain and Blake Brittain. 2024. Google Sued by US Artists over AI Image Generator. *Reuters* (April 2024).

[54] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models Are Few-Shot Learners. https://doi.org/10.48550/arXiv.2005.14165 arXiv:2005.14165 [cs]

[55] Adam Buick. 2024. Copyright and AI Training Data—Transparency to the Rescue? *Journal of Intellectual Property Law & Practice* (Dec. 2024), jpae102. https://doi.org/10.1093/jiplp/jpae102

[56] The Hindu Bureau. 2024. OpenAI CTO Dodges Questions around Training Data for Text-to-Video Generator Sora. *The Hindu* (March 2024).

[57] Carlos Caetano, Gabriel O. dos Santos, Caio Petrucci, Artur Barros, Camila Laranjeira, Leo Sampaio Ferraz Ribeiro, Júlia Fernandes de Mendonça, Jefersson A. dos Santos, and Sandra Avila. 2025. Neglected Risks: The Disturbing Reality of Children's Images in Datasets and the Urgent Call for Accountability. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency (FAccT '25)*. Association for Computing Machinery, New York, NY, USA, 2542–2553. https://doi.org/10.1145/3715275.3732166

[58] Alessandra Calvi, Gianclaudio Malgieri, and Dimitris Kotzinos. 2024. The Unfair Side of Privacy Enhancing Technologies: Addressing the Trade-Offs between PETs and Fairness. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency (FAccT '24)*. Association for Computing Machinery, New York, NY, USA, 2047–2059. https://doi.org/10.1145/3630106.3659024

[59] Mar Canet Sola and Varvara Guljajeva. 2024. Visions of Destruction: Exploring a Potential of Generative AI in Interactive Art. In *Proceedings of the 17th International Symposium on Visual Information Communication and Interaction (VINCI '24)*. Association for Computing Machinery, New York, NY, USA, 1–8. https://doi.org/10.1145/3678698.3687185

[60] Sarah H. Cen and Rohan Alur. 2024. From Transparency to Accountability and Back: A Discussion of Access and Evidence in AI Auditing. In *Proceedings of the 4th ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization (EAAMO '24)*. Association for Computing Machinery, New York, NY, USA, 1–14. https://doi.org/10.1145/3689904.3694711

[61] Kelvin Chan and Matt O'Brien. 2025. Getty Images and Stability AI Face off in British Copyright Trial That Will Test AI Industry. *AP News* (June 2025).

[62] Cheng Chen and S. Shyam Sundar. 2023. Is This AI Trained on Credible Data? The Effects of Labeling Quality and Performance Bias on User Trust. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23)*. Association for Computing Machinery, New York, NY, USA, 1–11. https://doi.org/10.1145/3544548.3580805

[63] Joséphus Cheung. 2023. GuanacoDataset (Revision 892e57a). https://doi.org/10.57967/hf/1423

[64] Madiha Zahrah Choksi, Ilan Mandel, and Sebastian Benthall. 2025. The Brief and Wondrous Life of Open Models. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency (FAccT '25)*. Association for Computing Machinery, New York, NY, USA, 3224–3240. https://doi.org/10.1145/3715275.3732206

[65] Danish Contractor, Daniel McDuff, Julia Katherine Haines, Jenny Lee, Christopher Hines, Brent Hecht, Nicholas Vincent, and Hanlin Li. 2022. Behavioral Use Licensing for Responsible AI. In *ACM International Conference Proceeding Series*. Association for Computing Machinery, 778–788. https://doi.org/10.1145/3531146.3533143 arXiv:2011.03116

[66] Eric Corbett and Emily Denton. 2023. Interrogating the T in FAccT. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency (FAccT '23)*. Association for Computing Machinery, New York, NY, USA, 1624–1634. https://doi.org/10.1145/3593013.3594104

[67] Kate Crawford and Trevor Paglen. 2021. Excavating AI: The Politics of Images in Machine Learning Training Sets. *AI & SOCIETY* 36, 4 (Dec. 2021), 1105–1116. https://doi.org/10.1007/s00146-021-01162-8

[68] Payton Croskey, Fabian Offert, Jennifer Jacobs, and Kai M. Thaler. 2025. Liberatory Collections and Ethical AI: Reimagining AI Development from Black Community Archives and Datasets. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency (FAccT '25)*. Association for Computing Machinery, New York, NY, USA, 900–913. https://doi.org/10.1145/3715275.3732058

[69] Alex Cukierman. 2009. The Limits of Transparency. *Economic Notes* 38, 1-2 (2009), 1–37. https://doi.org/10.1111/j.1468-0300.2009.00208.x

[70] Roxana Daneshjou, Mary P. Smith, Mary D. Sun, Veronica Rotemberg, and James Zou. 2021. Lack of Transparency and Potential Bias in Artificial Intelligence Data Sets and Algorithms: A Scoping Review. *JAMA Dermatology* 157, 11 (Nov. 2021), 1362–1369. https://doi.org/10.1001/jamadermatol.2021.3129

[71] Paul B. de Laat. 2018. Algorithmic Decision-Making Based on Machine Learning from Big Data: Can Transparency Restore Accountability? *Philosophy & Technology* 31, 4 (Dec. 2018), 525–541. https://doi.org/10.1007/s13347-017-0293-z

[72] Mark Díaz, Ian Kivlichan, Rachel Rosen, Dylan Baker, Razvan Amironesei, Vinodkumar Prabhakaran, and Emily Denton. 2022. CrowdWorkSheets: Accounting for Individual and Collective Identities Underlying Crowdsourced Dataset Annotation. In *2022 ACM Conference on Fairness, Accountability, and Transparency*. Association for Computing Machinery, New York, NY, USA, 2342–2351. https://doi.org/10.1145/3531146.3534647

[73] Jesse Dodge, Maarten Sap, Ana Marasović, William Agnew, Gabriel Ilharco, Dirk Groeneveld, Margaret Mitchell, and Matt Gardner. 2021. Documenting Large Webtext Corpora: A Case Study on the Colossal Clean Crawled Corpus. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih (Eds.). Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, 1286–1305. https://doi.org/10.18653/v1/2021.emnlp-main.98

[74] Andrés Domínguez Hernández, Shyam Krishna, Antonella Maia Perini, Michael Katell, SJ Bennett, Ann Borda, Youmna Hashem, Semeli Hadjiloizou, Sabeehah Mahomed, Smera Jayadeva, Mhairi Aitken, and David Leslie. 2024. Mapping the Individual, Social and Biospheric Impacts of Foundation Models. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency (FAccT '24)*. Association for Computing Machinery, New York, NY, USA, 776–796. https://doi.org/10.1145/3630106.3658939

[75] Upol Ehsan, Q. Vera Liao, Michael Muller, Mark O. Riedl, and Justin D. Weisz. 2021. Expanding Explainability: Towards Social Transparency in AI Systems. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems (CHI '21)*. Association for Computing Machinery, New York, NY, USA, 1–19. https://doi.org/10.1145/3411764.3445188

[76] Francisco Eiras, Aleksandar Petrov, Bertie Vidgen, Christian Schroeder, Fabio Pizzati, Katherine Elkins, Supratik Mukhopadhyay, Adel Bibi, Aaron Purewal, Csaba Botos, Fabro Steibel, Fazel Keshtkar, Fazl Barez, Genevieve Smith, Gianluca Guadagni, Jon Chun, Jordi Cabot, Joseph Imperial, Juan Arturo Nolazco, Lori Landay, Matthew Jackson, Phillip H. S. Torr, Trevor Darrell, Yong Lee, and Jakob Foerster. 2024. Risks and Opportunities of Open-Source Generative AI. https://doi.org/10.48550/arXiv.2405.08597 arXiv:2405.08597 [cs]

[77] Motahhare Eslami, Sarah Fox, Hong Shen, Bobbie Fan, Yu-Ru Lin, Rosta Farzan, and Beth Schwanke. 2025. From Margins to the Table: Charting the Potential for Public Participatory Governance of Algorithmic Decision Making. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency (FAccT '25)*. Association for Computing Machinery, New York, NY, USA, 2657–2670. https://doi.org/10.1145/3715275.3732173

[78] Florian Eyert and Paola Lopez. 2023. Rethinking Transparency as a Communicative Constellation. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency (FAccT '23)*. Association for Computing Machinery, New York, NY, USA, 444–454. https://doi.org/10.1145/3593013.3594010

[79] Heike Felzmann, Eduard Fosch Villaronga, Christoph Lutz, and Aurelia Tamò-Larrieux. 2019. Transparency You Can Trust: Transparency Requirements for Artificial Intelligence between Legal Norms and Contextual Concerns. *Big Data & Society* 6, 1 (Jan. 2019), 2053951719860542. https://doi.org/10.1177/2053951719860542

[80] Andrea Ferrario and Michele Loi. 2022. How Explainability Contributes to Trust in AI. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT '22)*. Association for Computing Machinery, New York, NY, USA, 1457–1466. https://doi.org/10.1145/3531146.3533202

[81] Matt Franchi, Hauke Sandhaus, Madiha Zahrah Choksi, Severin Engelmann, Wendy Ju, and Helen Nissenbaum. 2025. Privacy of Groups in Dense Street Imagery. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency (FAccT '25)*. Association for Computing Machinery, New York, NY, USA, 2874–2891. https://doi.org/10.1145/3715275.3732185

[82] Lynn J. Frewer. 2003. Trust, Transparency, and Social Context: Implications for Social Amplification of Risk. In *The Social Amplification of Risk*, Nick Pidgeon, Paul Slovic, and Roger E. Kasperson (Eds.). Cambridge University Press, Cambridge, 123–137. https://doi.org/10.1017/CBO9780511550461.006

[83] Batya Friedman. 1996. Value-Sensitive Design. *Interactions* 3, 6 (Dec. 1996), 16–23. https://doi.org/10.1145/242485.242493

[84] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé Iii, and Kate Crawford. 2021. Datasheets for Datasets. *Commun. ACM* 64, 12 (March 2021), 86–92. https://doi.org/10.1145/3458723 arXiv:1803.09010

[85] Edd Gent. 2024. The Tech Industry Can't Agree on What Open-Source AI Means. That's a Problem. *MIT Technology Review* (March 2024).

[86] Dr A. Shaji George, Dr T. Baskar, and Digvijay Pandey. 2024. Establishing Global AI Accountability: Training Data Transparency, Copyright, and Misinformation. *Partners Universal Innovative Research Publication* 2, 3 (June 2024), 75–91. https://doi.org/10.5281/zenodo.11659602

[87] Daniel Gillblad. 2023. Language Models for Everyone—Responsible and Transparent Development of Open Large Language Models. *Computer Sciences & Mathematics Forum* 8, 1 (2023), 51. https://doi.org/10.3390/cmsf2023008051

[88] Sharon Goldman. 2024. OpenAI's Sora: The Devil Is in the 'Details of the Data'. *VentureBeat* (March 2024).

[89] Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, Danny Wyatt, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Francisco Guzmán, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Govind Thattai, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jack Zhang, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Karthik Prasad, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Kushal Lakhotia, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Maria Tsimpoukelli, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Ning Zhang, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohan Maheswari, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vítor Albiero, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaofang Wang, Xiaoqing Ellen Tan, Xide Xia, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aayushi Srivastava, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Amos Teo, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Dong, Annie Franco, Anuj Goyal, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Ce Liu, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Cynthia Gao, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Eric-Tuan Le, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Filippos Kokkinos, Firat Ozgenel, Francesco Caggioni, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hakan Inan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter

Goldman, Hongyuan Zhan, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Ilias Leontiadis, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Janice Lam, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kiran Jagadeesh, Kun Huang, Kunal Chawla, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Miao Liu, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikhil Mehta, Nikolay Pavlovich Laptev, Ning Dong, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Rangaprabhu Parthasarathy, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Russ Howes, Ruty Rinott, Sachin Mehta, Sachin Siby, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Mahajan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shishir Patil, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Summer Deng, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Koehler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaojian Wu, Xiaolan Wang, Xilun Wu, Xinbo Gao, Yaniv Kleinman, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yu Zhao, Yuchen Hao, Yundi Qian, Yunlu Li, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, Zhiwei Zhao, and Zhiyu Ma. 2024. The Llama 3 Herd of Models. https://doi.org/10.48550/arXiv.2407.21783 arXiv:2407.21783 [cs]

[90] Colin M. Gray, Cristiana Teixeira Santos, Nataliia Bielova, and Thomas Mildner. 2024. An Ontology of Dark Patterns Knowledge: Foundations, Definitions, and a Pathway for Shared Knowledge-Building. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems (CHI '24)*. Association for Computing Machinery, New York, NY, USA, 1–22. https://doi.org/10.1145/3613904.3642436

[91] Stephan Grimmelikhuijsen. 2023. Explaining Why the Computer Says No: Algorithmic Transparency Affects the Perceived Trustworthiness of Automated Decision-Making. *Public Administration Review* 83, 2 (2023), 241–262. https://doi.org/10.1111/puar.13483

[92] Michael Gu, Ramasoumya Naraparaju, and Dongfang Zhao. 2024. Enhancing Data Provenance and Model Transparency in Federated Learning Systems – A Database Approach. https://doi.org/10.48550/arXiv.2403.01451 arXiv:2403.01451 [cs]

[93] Yandong Guo, Lei Zhang, Yuxiao Hu, Xiaodong He, and Jianfeng Gao. 2016. MS-celeb-1M: A Dataset and Benchmark for Large-Scale Face Recognition. In *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, Vol. 9907 LNCS. Springer Verlag, 87–102. https://doi.org/10.1007/978-3-319-46487-9_6 arXiv:1607.08221

[94] Ronan Hamon, Henrik Junklewitz, Gianclaudio Malgieri, Paul De Hert, Laurent Beslay, and Ignacio Sanchez. 2021. Impossible Explanations? Beyond Explainable AI in the GDPR from a COVID-19 Use Case Scenario. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT '21)*. Association for Computing Machinery, New York, NY, USA, 549–559. https://doi.org/10.1145/3442188.3445917

[95] Jack Hardinges, Elena Simperl, and Nigel Shadbolt. 2024. We Must Fix the Lack of Transparency Around the Data Used to Train Foundation Models. *Harvard Data Science Review* Special Issue 5 (May 2024). https://doi.org/10.1162/99608f92.a50ec6e6

[96] Philipp Hartmann and Joachim Henkel. 2020. The Rise of Corporate Science in AI: Data as a Strategic Resource. *Academy of Management Discoveries* 6, 3 (Sept. 2020), 359–381. https://doi.org/10.5465/amd.2019.0043

[97] Adam Harvey and Jules LaPlace. 2021. Exposing.Ai. https://exposing.ai/.

[98] Peter Henderson, Xuechen Li, Dan Jurafsky, Tatsunori Hashimoto, Mark A. Lemley, and Percy Liang. 2023. Foundation Models and Fair Use. https://doi.org/10.48550/arXiv.2303.15715 arXiv:2303.15715 [cs]

[99] Will Henshall. 2023. The Heated Debate Over Who Should Control Access to AI. https://time.com/6308604/meta-ai-access-open-source/.

[100] Michael Hind, Stephanie Houde, Jacquelyn Martino, Aleksandra Mojsilovic, David Piorkowski, John Richards, and Kush R. Varshney. 2020. Experiences with Improving the Transparency of AI Models and Services. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems (CHI EA '20)*. Association for Computing Machinery, New York, NY, USA, 1–8. https://doi.org/10.1145/3334480.3383051

[101] Dominik Hintersdorf, Lukas Struppek, and Kristian Kersting. 2025. Balancing Transparency and Risk: An Overview of the Security and Privacy Risks of Open-Source Machine Learning Models. In *Bridging the Gap Between AI and Reality*, Bernhard Steffen (Ed.). Springer Nature Switzerland, Cham, 269–283. https://doi.org/10.1007/978-3-031-73741-1_16

[102] Sarah Holland, Ahmed Hosny, Sarah Newman, Joshua Joseph, and Kasia Chmielinski. 2018. The Dataset Nutrition Label: A Framework To Drive Higher Data Quality Standards. (May 2018). arXiv:1805.03677

[103] Aspen Hopkins, Isabella Struckman, Kevin Klyman, and Susan S. Silbey. 2025. Recourse, Repair, Reparation, & Prevention: A Stakeholder Analysis of AI Supply Chains. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency (FAccT '25)*. Association for Computing Machinery, New York, NY, USA, 209–227. https://doi.org/10.1145/3715275.3732017

[104] John Howell and Stephanie Ifayemi. 2024. *Policy Alignment on AI Transparency: Analyzing Interoperability of Documentation Requirements across Eight Frameworks*. Technical Report. Partnership on AI.

[105] Saffron Huang and Divya Siddarth. 2023. Generative AI and the Digital Commons. https://doi.org/10.48550/arXiv.2303.11074 arXiv:2303.11074 [cs]

[106] Ben Hutchinson, Andrew Smart, Alex Hanna, Emily Denton, Christina Greer, Oddur Kjartansson, Parker Barnes, and Margaret Mitchell. 2021. Towards Accountability for Machine Learning Datasets: Practices from Software Engineering and Infrastructure. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT '21)*. Association for Computing Machinery, New York, NY, USA, 560–575. https://doi.org/10.1145/3442188.3445918

[107] Nanna Inie, Jeanette Falk, and Raghavendra Selvan. 2025. How CO2STLY Is CHI? The Carbon Footprint of Generative AI in HCI Research and What We Should Do About It. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI '25)*. Association for Computing Machinery, New York, NY, USA, 1–29. https://doi.org/10.1145/3706598.3714227

[108] Abigail Z. Jacobs and Hanna Wallach. 2021. Measurement and Fairness. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT '21)*. Association for Computing Machinery, New York, NY, USA, 375–385. https://doi.org/10.1145/3442188.3445901

[109] Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7B. https://doi.org/10.48550/arXiv.2310.06825 arXiv:2310.06825 [cs]

[110] Harry H. Jiang, Lauren Brown, Jessica Cheng, Mehtab Khan, Abhishek Gupta, Deja Workman, Alex Hanna, Johnathan Flowers, and Timnit Gebru. 2023. AI Art and Its Impact on Artists. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society (AIES '23)*. Association for Computing Machinery, New York, NY, USA, 363–374. https://doi.org/10.1145/3600211.3604681

[111] Amelia Jiménez-Sánchez, Natalia-Rozalia Avlona, Sarah de Boer, Víctor M. Campello, Aasa Feragen, Enzo Ferrante, Melanie Ganz, Judy Wawira Gichoya, Camila Gonzalez, Steff Groefsema, Alessa Hering, Adam Hulman, Leo Joskowicz, Dovile Juodelyte, Melih Kandemir, Thijs Kooi, Jorge del Pozo Lérida, Livie Yumeng Li, Andre Pacheco, Tim Rädsch, Mauricio Reyes, Théo Sourget, Bram van Ginneken, David Wen, Nina Weng, Jack Junchi Xu, Hubert Dariusz Zajaç, Maria A. Zuluaga, and Veronika Cheplygina. 2025. In the Picture: Medical Imaging Datasets, Artifacts, and Their Living Review. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency (FAccT '25)*. Association for Computing Machinery, New York, NY, USA, 511–531. https://doi.org/10.1145/3715275.3732035

[112] Nicola Jones. 2024. The AI Revolution Is Running out of Data. What Can Researchers Do? *Nature* 636, 8042 (Dec. 2024), 290–292. https://doi.org/10.1038/d41586-024-03990-2

[113] Kyu-Hwan Jung. 2023. Uncover This Tech Term: Foundation Model. *Korean Journal of Radiology* 24, 10 (Oct. 2023), 1038–1041. https://doi.org/10.3348/kjr.2023.0790

[114] Rie Kamikubo, Kyungjun Lee, and Hernisa Kacorri. 2023. Contributing to Accessibility Datasets: Reflections on Sharing Study Data by Blind People. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23)*. Association for Computing Machinery, New York, NY, USA, 1–18. https://doi.org/10.1145/3544548.3581337

[115] Jakko Kemper and Daan Kolkman. 2019. Transparent to Whom? No Algorithmic Accountability without a Critical Audience. *Information, Communication & Society* 22, 14 (Dec. 2019), 2081–2096. https://doi.org/10.1080/1369118X.2018.1477967

[116] Os Keyes, Jevan Hutson, and Meredith Durbin. 2019. A Mulching Proposal: Analysing and Improving an Algorithmic System for Turning the Elderly into High-Nutrient Slurry. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems (CHI EA '19)*. Association for Computing Machinery, New York, NY, USA, 1–11. https://doi.org/10.1145/3290607.3310433

[117] Moaiad Ahmad Khder. 2021. Web Scraping or Web Crawling: State of Art, Techniques, Approaches and Application. *International Journal of Advances in Soft Computing & Its Applications* 13, 3 (Nov. 2021), 144–168. https://doi.org/10.15849/ijasca.211128.11

[118] Chanwoo Kim, Soham U. Gadgil, Alex J. DeGrave, Zhuo Ran Cai, Roxana Daneshjou, and Su-In Lee. 2023. Fostering Transparent Medical Image AI via an Image-Text Foundation Model Grounded in Medical Literature. , 2023.06.07.23291119 pages. https://doi.org/10.1101/2023.06.07.23291119

[119] Chanwoo Kim, Soham U. Gadgil, Alex J. DeGrave, Jesutofunmi A. Omiye, Zhuo Ran Cai, Roxana Daneshjou, and Su-In Lee. 2024. Transparent Medical Image AI via an Image–Text Foundation Model Grounded in Medical Literature. *Nature Medicine* 30, 4 (April 2024), 1154–1165. https://doi.org/10.1038/s41591-024-02887-x

[120] Jennifer King, Daniel Ho, Arushi Gupta, Victor Wu, and Helen Webley-Brown. 2023. The Privacy-Bias Tradeoff: Data Minimization and Racial Disparity Assessments in U.S. Government. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency (FAccT '23)*. Association for Computing Machinery, New York, NY, USA, 492–505. https://doi.org/10.1145/3593013.3594015

[121] Kate Knibbs. 2024. Every AI Copyright Lawsuit in the US, Visualized. *Wired* (Dec. 2024).

[122] Bernard Koch, Emily Denton, Alex Hanna, and Jacob Gates Foster. 2021. Reduced, Reused and Recycled: The Life of a Dataset in Machine Learning Research. In *Thirty-Fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*.

[123] Jacob Leon Kröger, Milagros Miceli, and Florian Müller. 2021. How Data Can Be Used Against People: A Classification of Personal Data Misuses. *SSRN Electronic Journal* (Dec. 2021). https://doi.org/10.2139/SSRN.3887097

[124] Joshua Krook, Peter Winter, John Downer, and Jan Blockx. 2025. A Systematic Literature Review of Artificial Intelligence (AI) Transparency Laws in the European Union (EU) and United Kingdom (UK): A Socio-Legal Approach to AI Transparency Governance. *AI and Ethics* 5, 4 (Aug. 2025), 4069–4090. https://doi.org/10.1007/s43681-025-00674-z

[125] Augustin Landier and David Thesmar. 2011. Regulating Systemic Risk through Transparency: Tradeoffs in Making Data Public. https://doi.org/10.3386/w17664 national bureau of economic research:17664

[126] Anna Leschanowsky, Farnaz Salamatjoo, Zahra Kolagar, and Birgit Popp. 2025. Expert-Generated Privacy Q&A Dataset for Conversational AI and User Study Insights. In *Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems (CHI EA '25)*. Association for Computing Machinery, New York, NY, USA, 1–12. https://doi.org/10.1145/3706599.3720014

[127] Karen EC Levy and David Merritt Johns. 2016. When Open Data Is a Trojan Horse: The Weaponization of Transparency in Science and Governance. *Big Data & Society* 3, 1 (June 2016), 2053951715621568. https://doi.org/10.1177/2053951715621568

[128] Hanlin Li, Nicholas Vincent, Stevie Chancellor, and Brent Hecht. 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 (FAccT '23)*. Association for Computing Machinery, New York, NY, USA, 1151–1161. https://doi.org/10.1145/3593013.3594070

[129] Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, Benjamin Newman, Binhang Yuan, Bobby Yan, Ce Zhang, Christian Cosgrove, Christopher D. Manning, Christopher Ré, Diana Acosta-Navas, Drew A. Hudson, Eric Zelikman, Esin Durmus, Faisal Ladhak, Frieda Rong, Hongyu Ren, Huaxiu Yao, Jue Wang, Keshav Santhanam, Laurel Orr, Lucia Zheng, Mert Yuksekgonul, Mirac Suzgun, Nathan Kim, Neel Guha, Niladri Chatterji, Omar Khattab, Peter Henderson, Qian Huang, Ryan Chi, Sang Michael Xie, Shibani Santurkar, Surya Ganguli, Tatsunori Hashimoto, Thomas Icard, Tianyi Zhang, Vishrav Chaudhary, William Wang, Xuechen Li, Yifan Mai, Yuhui Zhang, and Yuta Koreeda. 2023. Holistic Evaluation of Language Models. https://doi.org/10.48550/arXiv.2211.09110 arXiv:2211.09110 [cs]

[130] Q. Vera Liao and Jennifer Wortman Vaughan. 2023. AI Transparency in the Age of LLMs: A Human-Centered Research Roadmap. (June 2023). arXiv:2306.01941

[131] Andreas Liesenfeld, Alianda Lopez, and Mark Dingemanse. 2023. Opening up ChatGPT: Tracking Openness, Transparency, and Accountability in Instruction-Tuned Text Generators. In *Proceedings of the 5th International Conference on Conversational User Interfaces (CUI '23)*. Association for Computing Machinery, New York, NY, USA, 1–6. https://doi.org/10.1145/3571884.3604316

[132] Sebastian Linxen, Christian Sturm, Florian Brühlmann, Vincent Cassau, Klaus Opwis, and Katharina Reinecke. 2021. How WEIRD Is CHI?. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems (CHI '21)*. Association for Computing Machinery, New York, NY, USA, 1–14. https://doi.org/10.1145/3411764.3445488

[133] Zhengzhong Liu, Aurick Qiao, Willie Neiswanger, Hongyi Wang, Bowen Tan, Tianhua Tao, Junbo Li, Yuqi Wang, Suqi Sun, Omkar Pangarkar, Richard Fan, Yi Gu, Victor Miller, Yonghao Zhuang, Guowei He, Haonan Li, Fajri Koto, Liping Tang, Nikhil Ranjan, Zhiqiang Shen, Roberto Iriondo, Cun Mu, Zhiting Hu, Mark Schulze, Preslav Nakov, Timothy Baldwin, and Eric P. Xing. 2024. LLM360: Towards Fully Transparent Open-Source LLMs. In *First Conference on Language Modeling*.

[134] Shayne Longpre, Robert Mahari, Anthony Chen, Naana Obeng-Marnu, Damien Sileo, William Brannon, Niklas Muennighoff, Nathan Khazam, Jad Kabbara, Kartik Perisetla, Xinyi Wu, Enrico Shippole, Kurt Bollacker, Tongshuang Wu, Luis Villa, Sandy Pentland, and Sara Hooker. 2023. The Data Provenance Initiative: A Large Scale Audit of Dataset Licensing & Attribution in AI. https://doi.org/10.48550/arXiv.2310.16787 arXiv:2310.16787 [cs]

[135] Shayne Longpre, Robert Mahari, Ariel Lee, Campbell Lund, Hamidah Oderinwale, William Brannon, Nayan Saxena, Naana Obeng-Marnu, Tobin South, Cole Hunter, Christopher Klamm, Hailey Schoelkopf, Nikhil Singh, Manuel Cherep, Mustafa Anis, An Dinh, Caroline Chitongo, Da Yin, Damien Sileo, Deividas Mataciunas, Diganta Misra, Emad Alghamdi, Enrico Shippole, Jianguo Zhang, Joanna Materzynska, Kun Qian, Kush Tiwary, Lester Miranda, Manan Dey, Minnie Liang, Niklas Muennighoff, Seonghyeon Ye, Seungone Kim, Shrestha Mohanty, Vivek Sharma, Vu Minh Chien, Xuhui Zhou, Yizhi Li, Caiming Xiong, Luis Villa, Stella Biderman, Hanlin Li, Daphne Ippolito, Sara Hooker, and Jad Kabbara. 2024. Consent in Crisis: The Rapid Decline of the AI Data Commons. (2024).

[136] Shayne Longpre, Robert Mahari, Naana Obeng-Marnu, William Brannon, Tobin South, Jad Kabbara, and Sandy Pentland. 2024. Data Authenticity, Consent, and Provenance for AI Are All Broken: What Will It Take to Fix Them? *An MIT Exploration of Generative AI* (March 2024). https://doi.org/10.21428/e4baedd9.a650f77d

[137] Juniper Lovato, Julia Witte Zimmerman, Isabelle Smith, Peter Dodds, and Jennifer L. Karson. 2024. Foregrounding Artist Opinions: A Survey Study on Transparency, Ownership, and Fairness in AI Generative Art. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society* 7 (Oct. 2024), 905–916. https://doi.org/10.1609/aies.v7i1.31691

[138] Sylvia Lu. 2020–2021. Algorithmic Opacity, Private Accountability, and Corporate Social Disclosure in the Age of Artificial Intelligence. *Vanderbilt Journal of Entertainment & Technology Law* 23 (2020–2021), 99.

[139] Heiko Ludwig, Yi Zhou, Syed Zawad, Yuya Ong, Pengyuan Li, Eric Butler, and Eelaaf Zahid. 2024. Towards Collecting Royalties for Copyrighted Data for Generative Models. In *2024 IEEE International Conference on Web Services (ICWS)*. 20–26. https://doi.org/10.1109/ICWS62655.2024.00020

[140] Angela Luna. 2024. Open-Source AI: The Debate That Could Redefine AI Innovation.

[141] Rongjun Ma, Caterina Maidhof, Juan Carlos Carrillo, Janne Lindqvist, and Jose Such. 2025. Privacy Perceptions of Custom GPTs by Users and Creators. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI '25)*. Association for Computing Machinery, New York, NY, USA, 1–18. https://doi.org/10.1145/3706598.3713540

[142] Abdul Majeed and Seong Oun Hwang. 2023. When AI Meets Information Privacy: The Adversarial Role of AI in Data Sharing Scenario. *IEEE Access* 11 (2023), 76177–76195. https://doi.org/10.1109/ACCESS.2023.3297646

[143] Mariavittoria Masotina, Elena Musi, and Anna Spagnolli. 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 (CHItaly '23)*. Association for Computing Machinery, New York, NY, USA, 1–8. https://doi.org/10.1145/3605390.3605413

[144] Nora McDonald, Sarita Schoenebeck, and Andrea Forte. 2019. Reliability and Inter-Rater Reliability in Qualitative Research: Norms and Guidelines for CSCW and HCI Practice. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 23. https://doi.org/10.1145/3359174

[145] Scott Mayer McKinney, Alan Karthikesalingam, Daniel Tse, Christopher J. Kelly, Yun Liu, Greg S. Corrado, and Shravya Shetty. 2020. Reply to: Transparency and Reproducibility in Artificial Intelligence. *Nature* 586, 7829 (Oct. 2020), E17–E18. https://doi.org/10.1038/s41586-020-2767-x

[146] Angelina McMillan-Major, Emily M. Bender, and Batya Friedman. 2024. Data Statements: From Technical Concept to Community Practice. *ACM Journal on Responsible Computing* 1, 1 (March 2024), 1–17. https://doi.org/10.1145/3594737

[147] Afshin Mehrpouya and Marie-Laure Djelic. 2014. Transparency: From Enlightenment to Neoliberalism or When a Norm of Liberation Becomes a Tool of Governing. https://doi.org/10.2139/ssrn.2499402 social science research network:2499402

[148] Michele Merler, Nalini Ratha, Rogerio S. Feris, and John R. Smith. 2019. Diversity in Faces. (Jan. 2019). arXiv:1901.10436

[149] Cade Metz, Cecilia Kang, Sheera Frenkel, Stuart A. Thompson, and Nico Grant. 2024. How Tech Giants Cut Corners to Harvest Data for A.I. *The New York Times* (April 2024).

[150] Milagros Miceli, Tianling Yang, Laurens Naudts, Martin Schuessler, Diana Serbanescu, and Alex Hanna. 2021. Documenting Computer Vision Datasets: An Invitation to Reflexive Data Practices. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT '21)*. ACM, New York, NY, USA, 161–172. https://doi.org/10.1145/3442188.3445880

[151] Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. 2018. Model Cards for Model Reporting. In *Proceedings of the 2019 Conference on Fairness, Accountability, and Transparency*. Association for Computing Machinery, Inc, 220–229. https://doi.org/10.1145/3287560.3287596 arXiv:1810.03993v2

[152] J. W. Moses and T. L. Knutsen. 2019. *Ways of Knowing: Competing Methodologies in Social and Political Research*. Bloomsbury Publishing.

[153] Devesh Narayanan. 2023. Welfarist Moral Grounding for Transparent AI. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency (FAccT '23)*. Association for Computing Machinery, New York, NY, USA, 64–76. https://doi.org/10.1145/3593013.3593977

[154] National Institute of Standards and Technology (US). 2024. *Artificial Intelligence Risk Management Framework : Generative Artificial Intelligence Profile*. Technical Report error: 600-1. National Institute of Standards and Technology (U.S.), Gaithersburg, MD. error: 600–1 pages. https://doi.org/10.6028/NIST.AI.600-1

[155] C. Thi Nguyen. 2022. Transparency Is Surveillance. *Philosophy and Phenomenological Research* 105, 2 (2022), 331–361. https://doi.org/10.1111/phpr.12823

[156] Matthew J. Page, Joanne E. McKenzie, Patrick M. Bossuyt, Isabelle Boutron, Tammy C. Hoffmann, Cynthia D. Mulrow, Larissa Shamseer, Jennifer M. Tetzlaff, Elie A. Akl, Sue E. Brennan, Roger Chou, Julie Glanville, Jeremy M. Grimshaw, Asbjørn Hróbjartsson, Manoj M. Lalu, Tianjing Li, Elizabeth W. Loder, Evan Mayo-Wilson, Steve McDonald, Luke A. McGuinness, Lesley A. Stewart, James Thomas, Andrea C. Tricco, Vivian A. Welch, Penny Whiting, and David Moher. 2021. The PRISMA 2020 Statement: An Updated Guideline for Reporting Systematic Reviews. *BMJ* 372 (March 2021), n71. https://doi.org/10.1136/bmj.n71

[157] Lawrence A. Palinkas, Sarah M. Horwitz, Carla A. Green, Jennifer P. Wisdom, Naihua Duan, and Kimberly Hoagwood. 2015. Purposeful Sampling for Qualitative Data Collection and Analysis in Mixed Method Implementation Research. *Administration and policy in mental health* 42, 5 (Sept. 2015), 533–544. https://doi.org/10.1007/s10488-013-0528-y

[158] Orestis Papakyriakopoulos, Anna Seo Gyeong Choi, William Thong, Dora Zhao, Jerone Andrews, Rebecca Bourke, Alice Xiang, and Allison Koenecke. 2023. Augmented Datasheets for Speech Datasets and Ethical Decision-Making. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency (FAccT '23)*. Association for Computing Machinery, New York, NY, USA, 881–904. https://doi.org/10.1145/3593013.3594049

[159] Beata Paragi. 2024. The Politics of Opacity and Transparency in Non-European Contexts. In *Screening by International Aid Organizations Operating in the Global South: Mitigating Risks of Generosity*, Beata Paragi (Ed.). Springer Nature Switzerland, Cham, 133–173. https://doi.org/10.1007/978-3-031-54165-0_5

[160] Tamara Paris, Ajung Moon, and Jin L.C. Guo. 2025. Opening the Scope of Openness in AI. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency (FAccT '25)*. Association for Computing Machinery, New York, NY, USA, 1293–1311. https://doi.org/10.1145/3715275.3732087

[161] Irene V. Pasquetto, Zoë Cullen, Andrea Thomer, and Morgan Wofford. 2024. What Is Research Data "Misuse"? And How Can It Be Prevented or Mitigated? *Journal of the Association for Information Science and Technology* 75, 12 (2024), 1413–1429. https://doi.org/10.1002/asi.24944

[162] Federica Pepe, Vittoria Nardone, Antonio Mastropaolo, Gabriele Bavota, Gerardo Canfora, and Massimiliano Di Penta. 2024. How Do Hugging Face Models Document Datasets, Bias, and Licenses? An Empirical Study. In *Proceedings of the 32nd IEEE/ACM International Conference on Program Comprehension (ICPC '24)*. Association for Computing Machinery, New York, NY, USA, 370–381. https://doi.org/10.1145/3643916.3644412

[163] Aleksandra Piktus, Christopher Akiki, Paulo Villegas, Hugo Laurençon, Gérard Dupont, Sasha Luccioni, Yacine Jernite, and Anna Rogers. 2023. The ROOTS Search Tool: Data Transparency for LLMs. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, Danushka Bollegala, Ruihong Huang, and Alan Ritter (Eds.). Association for Computational Linguistics, Toronto, Canada, 304–314. https://doi.org/10.18653/v1/2023.acl-demo.29

[164] Lindsay Poirier, Juniper Huang, and Casey MacGibbon. 2025. What Remains Opaque in Transparency Initiatives: Visualizing Phantom Reductions through Devious Data Analysis. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency (FAccT '25)*. Association for Computing Machinery, New York, NY, USA, 1703–1714. https://doi.org/10.1145/3715275.3732114

[165] Mahima Pushkarna, Andrew Zaldivar, and Oddur Kjartansson. 2022. Data Cards: Purposeful and Transparent Dataset Documentation for Responsible AI. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT '22)*. Association for Computing Machinery, New York, NY, USA, 1776–1826. https://doi.org/10.1145/3531146.3533231

[166] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language Models Are Unsupervised Multitask Learners. (2019).

[167] Divya Ramesh, Vaishnav Kameswaran, Ding Wang, and Nithya Sambasivan. 2022. How Platform-User Power Relations Shape Algorithmic Accountability: A Case Study of Instant Loan Platforms and Financially Stressed Users in India. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT '22)*. Association for Computing Machinery, New York, NY, USA, 1917–1928. https://doi.org/10.1145/3531146.3533237

[168] Karoline Reinhardt. 2023. Trust and Trustworthiness in AI Ethics. *AI and Ethics* 3, 3 (Aug. 2023), 735–744. https://doi.org/10.1007/s43681-022-00200-5

[169] Donald S. Rep. Beyer. 2023. H.R.6881 - 118th Congress (2023-2024): AI Foundation Model Transparency Act of 2023.

[170] Graham R.Gibbs. 2007. *Analyzing Qualitative Data*. SAGE Publications, Ltd. https://doi.org/10.4135/9781849208574

[171] Graham R.Gibbs. 2007. Thematic Coding and Categorizing. In *Analyzing Qualitative Data*. SAGE Publications, Ltd, 38–55. https://doi.org/10.4135/9781849208574

[172] Kylie Robison. 2024. Open-Source AI Must Reveal Its Training Data, per New OSI Definition. https://www.theverge.com/2024/10/28/24281820/open-source-initiative-definition-artificial-intelligence-meta-llama.

[173] Juan Manuel Corchado Rodriguez, Karen Mossberger, Pauline Hope Cheong, Rita Yi Man Li, and Tan Yigitcanlar. 2024. Local Governments Are Using AI without Clear Rules or Policies, and the Public Has No Idea. *The Conversation* (Dec. 2024).

[174] Daniel Rodriguez Maffioli. 2023. Copyright in Generative AI Training: Balancing Fair Use through Standardization and Transparency. https://doi.org/10.2139/ssrn.4579322 social science research network:4579322

[175] Negar Rostamzadeh, Diana Mincu, Subhrajit Roy, Andrew Smart, Lauren Wilcox, Mahima Pushkarna, Jessica Schrouff, Razvan Amironesei, Nyalleng Moorosi, and Katherine Heller. 2022. Healthsheet: Development of a Transparency Artifact for Health Datasets. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT '22)*. Association for Computing Machinery, New York, NY, USA, 1943–1961. https://doi.org/10.1145/3531146.3533239

[176] Malak Sadek, Rafael A. Calvo, and Céline Mougenot. 2024. Designing Value-Sensitive AI: A Critical Review and Recommendations for Socio-Technical Design Processes. *AI and Ethics* 4, 4 (Nov. 2024), 949–967. https://doi.org/10.1007/s43681-023-00373-7

[177] Christopher James Sampson, Renée Arnold, Stirling Bryan, Philip Clarke, Sean Ekins, Anthony Hatswell, Neil Hawkins, Sue Langham, Deborah Marshall, Mohsen Sadatsafavi, Will Sullivan, Edward C. F. Wilson, and Tim Wrightson. 2019. Transparency in Decision Modelling: What, Why, Who and How? *PharmacoEconomics* 37, 11 (Nov. 2019), 1355–1369. https://doi.org/10.1007/s40273-019-00819-z

[178] Sheeba Samuel, Frank Löffler, and Birgitta König-Ries. 2021. Machine Learning Pipelines: Provenance, Reproducibility and FAIR Data Principles. In *Provenance and Annotation of Data and Processes*, Boris Glavic, Vanessa Braganholo, and David Koop (Eds.). Springer International Publishing, Cham, 226–230. https://doi.org/10.1007/978-3-030-80960-7_17

[179] Cristian Santesteban and Shayne Longpre. 2020. How Big Data Confers Market Power to Big Tech: Leveraging the Perspective of Data Science. *The Antitrust Bulletin* 65, 3 (Sept. 2020), 459–485. https://doi.org/10.1177/0003603X20934212

[180] Devansh Saxena, Karla Badillo-Urquiola, Pamela J. Wisniewski, and Shion Guha. 2020. A Human-Centered Review of Algorithms Used within the U.S. Child Welfare System. *Conference on Human Factors in Computing Systems - Proceedings* (April 2020). https://doi.org/10.1145/3313831.3376229 arXiv:2003.03541

[181] Nicolas Scharowski, Michaela Benk, Swen J. Kühne, Léane Wettstein, and Florian Brühlmann. 2023. Certification Labels for Trustworthy AI: Insights From an Empirical Mixed-Method Study. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency (FAccT '23)*. Association for Computing Machinery, New York, NY, USA, 248–260. https://doi.org/10.1145/3593013.3593994

[182] Morgan Klaus Scheuerman. 2024. In the Walled Garden: Challenges and Opportunities for Research on the Practices of the AI Tech Industry. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency (FAccT '24)*. Association for Computing Machinery, New York, NY, USA, 456–466. https://doi.org/10.1145/3630106.3658918

[183] Morgan Klaus Scheuerman, Katy Weathington, Tarun Mugunthan, Emily Denton, and Casey Fiesler. 2023. From Human to Data to Dataset: Mapping the Traceability of Human Subjects in Computer Vision Datasets. *Proc. ACM Hum.-Comput. Interact.* 7, CSCW1 (April 2023). https://doi.org/10.1145/3579488

[184] Elizabeth Seger, Noemi Dreksler, Richard Moulange, Emily Dardaman, Jonas Schuett, K. Wei, Christoph Winter, Mackenzie Arnold, Seán Ó hÉigeartaigh, Anton Korinek, Markus Anderljung, Ben Bucknall, Alan Chan, Eoghan Stafford, Leonie Koessler, Aviv Ovadya, Ben Garfinkel, Emma Bluemke, Michael Aird, Patrick Levermore, Julian Hazell, and Abhishek Gupta. 2023. Open-Sourcing Highly Capable Foundation Models: An Evaluation of Risks, Benefits, and Alternative Methods for Pursuing Open-Source Objectives. *SSRN Electronic Journal* (2023). https://doi.org/10.2139/ssrn.4596436

[185] Ali Akbar Septiandri, Marios Constantinides, Mohammad Tahaei, and Daniele Quercia. 2023. WEIRD FAccTs: How Western, Educated, Industrialized, Rich,

and Democratic Is FAccT?. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency (FAccT '23)*. Association for Computing Machinery, New York, NY, USA, 160–171. https://doi.org/10.1145/3593013.3593985

[186] Orit Shaer, Angelora Cooper, Osnat Mokryn, Andrew L Kun, and Hagit Ben Shoshan. 2024. AI-Augmented Brainwriting: Investigating the Use of LLMs in Group Ideation. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems (CHI '24)*. Association for Computing Machinery, New York, NY, USA, 1–17. https://doi.org/10.1145/3613904.3642414

[187] Yashothara Shanmugarasa, Shidong Pan, Ming Ding, Dehai Zhao, and Thierry Rakotoarivelo. 2025. Privacy Meets Explainability: Managing Confidential Data and Transparency Policies in LLM-Empowered Science. In *Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems (CHI EA '25)*. Association for Computing Machinery, New York, NY, USA, 1–8. https://doi.org/10.1145/3706599.3720099

[188] Renee Shelby, Shalaleh Rismani, and Negar Rostamzadeh. 2024. Generative AI in Creative Practice: ML-Artist Folk Theories of T2I Use, Harm, and Harm-Reduction. In *Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI '24)*. Association for Computing Machinery, New York, NY, USA, 1–17. https://doi.org/10.1145/3613904.3642461

[189] Congzhen Shi, Ryan Rezai, Jiaxi Yang, Qi Dou, and Xiaoxiao Li. 2024. A Survey on Trustworthiness in Foundation Models for Medical Image Analysis. https://doi.org/10.48550/arXiv.2407.15851 arXiv:2407.15851 [cs]

[190] Meixue Si, Shidong Pan, Dianshu Liao, Xiaoyu Sun, Zhen Tao, Wenchang Shi, and Zhenchang Xing. 2024. A Solution toward Transparent and Practical AI Regulation: Privacy Nutrition Labels for Open-source Generative AI-based Applications. https://doi.org/10.48550/arXiv.2407.15407 arXiv:2407.15407 [cs]

[191] Huzaifa Sidhpurwala, Garth Mollett, Emily Fox, Mark Bestavros, and Huamin Chen. 2024. Building Trust: Foundations of Security, Safety and Transparency in AI. https://doi.org/10.48550/arXiv.2411.12275 arXiv:2411.12275 [cs]

[192] David A. Siegel. 2003. The Business Case for User-Centered Design: Increasing Your Power of Persuasion. *interactions* 10, 3 (May 2003), 30–36. https://doi.org/10.1145/769759.769772

[193] Isabella Barbosa Silva, Elsa Oliveira, Ricardo Melo, Luís Rosado, César Gálvez-Barrón, Irene Bernadet Heijink, Sem Hoogteijling, and Iñigo Gabilondo. 2025. Designing for Qualitative Evaluation of Synthetic Medical Data. In *Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems (CHI EA '25)*. Association for Computing Machinery, New York, NY, USA, 1–11. https://doi.org/10.1145/3706599.3720274

[194] Jan Simson, Alessandro Fabris, and Christoph Kern. 2024. Lazy Data Practices Harm Fairness Research. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency (FAccT '24)*. Association for Computing Machinery, New York, NY, USA, 642–659. https://doi.org/10.1145/3630106.3658931

[195] Robert Soden, Austin Toombs, and Michaelanne Thomas. 2024. Evaluating Interpretive Research in HCI. *interactions* 31, 1 (Jan. 2024), 38–42. https://doi.org/10.1145/3633200

[196] Luca Soldaini, Rodney Kinney, Akshita Bhagia, Dustin Schwenk, David Atkinson, Russell Authur, Ben Bogin, Khyathi Chandu, Jennifer Dumas, Yanai Elazar, Valentin Hofmann, Ananya Jha, Sachin Kumar, Li Lucy, Xinxi Lyu, Nathan Lambert, Ian Magnusson, Jacob Morrison, Niklas Muennighoff, Aakanksha Naik, Crystal Nam, Matthew Peters, Abhilasha Ravichander, Kyle Richardson, Zejiang Shen, Emma Strubell, Nishant Subramani, Oyvind Tafjord, Evan Walsh, Luke Zettlemoyer, Noah Smith, Hannaneh Hajishirzi, Iz Beltagy, Dirk Groeneveld, Jesse Dodge, and Kyle Lo. 2024. Dolma: An Open Corpus of Three Trillion Tokens for Language Model Pretraining Research. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Lun-Wei Ku, Andre Martins, and Vivek Srikumar (Eds.). Association for Computational Linguistics, Bangkok, Thailand, 15725–15788. https://doi.org/10.18653/v1/2024.acl-long.840

[197] Aruni Sony. 2024. OpenAI, Authors Get Rules Set for Inspecting AI Training Data. *Bloomberg Law* (Sept. 2024).

[198] Manjunath Srinivas, S Vamsi Krishna Reddy, Manoj N M, and Hatsuho Miyazawa. 2024. Evaluation of ChatGPT, Gemini and Llama-2 for E-commerce Product Attribute Extraction. In *Proceedings of the 2024 10th International Conference on E-Society, e-Learning and e-Technologies (ICSLT) (ICSLT '24)*. Association for Computing Machinery, New York, NY, USA, 43–48. https://doi.org/10.1145/3678610.3678619

[199] Eugenia Stamboliev. 2023. Proposing a Postcritical AI Literacy: Why We Should Worry Less about Algorithmic Transparency and More about Citizen Empowerment. *Media Theory* 7, 1 (Sept. 2023), 202–232.

[200] Jonathan Stempel. 2023. NY Times Sues OpenAI, Microsoft for Infringing Copyrighted Works. *Reuters* (Dec. 2023).

[201] Remy Stewart. 2021. Big Data and Belmont: On the Ethics and Research Implications of Consumer-Based Datasets. *Big Data & Society* 8, 2 (July 2021), 20539517211048183. https://doi.org/10.1177/20539517211048183

[202] Artur Strzelecki. 2024. To Use or Not to Use ChatGPT in Higher Education? A Study of Students' Acceptance and Use of Technology. *Interactive Learning Environments* 32, 9 (Oct. 2024), 5142–5155. https://doi.org/10.1080/10494820.2023.2209881

[203] Harini Suresh, Emily Tseng, Meg Young, Mary Gray, Emma Pierson, and Karen Levy. 2024. Participation in the Age of Foundation Models. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency (FAccT '24)*. Association for Computing Machinery, New York, NY, USA, 1609–1621. https://doi.org/10.1145/3630106.3658992

[204] Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, Soroosh Mariooryad, Yifan Ding, Xinyang Geng, Fred Alcober, Roy Frostig, Mark Omernick, Lexi Walker, Cosmin Paduraru, Christina Sorokin, Andrea Tacchetti, Colin Gaffney, Samira Daruki, Olcan Sercinoglu, Zach Gleicher, Juliette Love, Paul Voigtlaender, Rohan Jain, Gabriela Surita, Kareem Mohamed, Rory Blevins, Junwhan Ahn, Tao Zhu, Kornraphop Kawintiranon, Orhan Firat, Yiming Gu, Yujing Zhang, Matthew Rahtz, Manaal Faruqui, Natalie Clay, Justin Gilmer, J. D. Co-Reyes, Ivo Penchev, Rui Zhu, Nobuyuki Morioka, Kevin Hui, Krishna Haridasan, Victor Campos, Mahdis Mahdieh, Mandy Guo, Samer Hassan, Kevin Kilgour, Arpi Vezer, Heng-Tze Cheng, Raoul de Liedekerke, Siddharth Goyal, Paul Barham, D. J. Strouse, Seb Noury, Jonas Adler, Mukund Sundararajan, Sharad Vikram, Dmitry Lepikhin, Michela Paganini, Xavier Garcia, Fan Yang, Dasha Valter, Maja Trebacz, Kiran Vodrahalli, Chulayuth Asawaroengchai, Roman Ring, Norbert Kalb, Livio Baldini Soares, Siddhartha Brahma, David Steiner, Tianhe Yu, Fabian Mentzer, Antoine He, Lucas Gonzalez, Bibo Xu, Raphael Lopez Kaufman, Laurent El Shafey, Junhyuk Oh, Tom Hennigan, George van den Driessche, Seth Odoom, Mario Lucic, Becca Roelofs, Sid Lall, Amit Marathe, Betty Chan, Santiago Ontanon, Luheng He, Denis Teplyashin, Jonathan Lai, Phil Crone, Bogdan Damoc, Lewis Ho, Sebastian Riedel, Karel Lenc, Chih-Kuan Yeh, Aakanksha Chowdhery, Yang Xu, Mehran Kazemi, Ehsan Amid, Anastasia Petrushkina, Kevin Swersky, Ali Khodaei, Gowoon Chen, Chris Larkin, Mario Pinto, Geng Yan, Adria Puigdomenech Badia, Piyush Patil, Steven Hansen, Dave Orr, Sebastien M. R. Arnold, Jordan Grimstad, Andrew Dai, Sholto Douglas, Rishika Sinha, Vikas Yadav, Xi Chen, Elena Gribovskaya, Jacob Austin, Jeffrey Zhao, Kaushal Patel, Paul Komarek, Sophia Austin, Sebastian Borgeaud, Linda Friso, Abhimanyu Goyal, Ben Caine, Kris Cao, Da-Woon Chung, Matthew Lamm, Gabe Barth-Maron, Thais Kagohara, Kate Olszewska, Mia Chen, Kaushik Shivakumar, Rishabh Agarwal, Harshal Godhia, Ravi Rajwar, Javier Snaider, Xerxes Dotiwalla, Yuan Liu, Aditya Barua, Victor Ungureanu, Yuan Zhang, Bat-Orgil Batsaikhan, Mateo Wirth, James Qin, Ivo Danihelka, Tulsee Doshi, Martin Chadwick, Jilin Chen, Sanil Jain, Quoc Le, Arjun Kar, Madhu Gurumurthy, Cheng Li, Ruoxin Sang, Fangyu Liu, Lampros Lamprou, Rich Munoz, Nathan Lintz, Harsh Mehta, Heidi Howard, Malcolm Reynolds, Lora Aroyo, Quan Wang, Lorenzo Blanco, Albin Cassirer, Jordan Griffith, Dipanjan Das, Stephan Lee, Jakub Sygnowski, Zach Fisher, James Besley, Richard Powell, Zafarali Ahmed, Dominik Paulus, David Reitter, Zalan Borsos, Rishabh Joshi, Aedan Pope, Steven Hand, Vittorio Selo, Vihan Jain, Nikhil Sethi, Megha Goel, Takaki Makino, Rhys May, Zhen Yang, Johan Schalkwyk, Christina Butterfield, Anja Hauth, Alex Goldin, Will Hawkins, Evan Senter, Sergey Brin, Oliver Woodman, Marvin Ritter, Eric Noland, Minh Giang, Vijay Bolina, Lisa Lee, Tim Blyth, Ian Mackinnon, Machel Reid, Obaid Sarvana, David Silver, Alexander Chen, Lily Wang, Loren Maggiore, Oscar Chang, Nithya Attaluri, Gregory Thornton, Chung-Cheng Chiu, Oskar Bunyan, Nir Levine, Timothy Chung, Evgenii Eltyshev, Xiance Si, Timothy Lillicrap, Demetra Brady, Vaibhav Aggarwal, Boxi Wu, Yuanzhong Xu, Ross McIlroy, Kartikeya Badola, Paramjit Sandhu, Erica Moreira, Wojciech Stokowiec, Ross Hemsley, Dong Li, Alex Tudor, Pranav Shyam, Elahe Rahimtoroghi, Salem Haykal, Pablo Sprechmann, Xiang Zhou, Diana Mincu, Yujia Li, Ravi Addanki, Kalpesh Krishna, Xiao Wu, Alexandre Frechette, Matan Eyal, Allan Dafoe, Dave Lacey, Jay Whang, Thi Avrahami, Ye Zhang, Emanuel Taropa, Hanzhao Lin, Daniel Toyama, Eliza Rutherford, Motoki Sano, HyunJeong Choe, Alex Tomala, Chalence Safranek-Shrader, Nora Kassner, Mantas Pajarskas, Matt Harvey, Sean Sechrist, Meire Fortunato, Christina Lyu, Gamaleldin Elsayed, Chenkai Kuang, James Lottes, Eric Chu, Chao Jia, Chih-Wei Chen, Peter Humphreys, Kate Baumli, Connie Tao, Rajkumar Samuel, Cicero Nogueira dos Santos, Anders Andreassen, Nemanja Rakićević, Dominik Grewe, Aviral Kumar, Stephanie Winkler, Jonathan Caton, Andrew Brock, Sid Dalmia, Hannah Sheahan, Iain Barr, Yingjie Miao, Paul Natsev, Jacob Devlin, Feryal Behbahani, Flavien Prost, Yanhua Sun, Artiom Myaskovsky, Thanumalayan Sankaranarayana Pillai, Dan Hurt, Angeliki Lazaridou, Xi Xiong, Ce Zheng, Fabio Pardo, Xiaowei Li, Dan Horgan, Joe Stanton, Moran Ambar, Fei Xia, Alejandro Lince, Mingqiu Wang, Basil Mustafa, Albert Webson, Hyo Lee, Rohan Anil, Martin Wicke, Timothy Dozat, Abhishek Sinha, Enrique Piqueras, Elahe Dabir, Shyam Upadhyay, Anudhyan Boral, Lisa Anne Hendricks, Corey Fry, Josip Djolonga, Yi Su, Jake Walker, Jane Labanowski, Ronny Huang, Vedant Misra, Jeremy Chen, R. J. Skerry-Ryan, Avi Singh, Shruti Rijhwani, Dian Yu, Alex Castro-Ros, Beer Changpinyo, Romina Datta, Sumit Bagri, Arnar Mar Hrafnkelsson, Marcello Maggioni, Daniel Zheng, Yury Sulsky, Shaobo Hou, Tom Le Paine, Antoine Yang, Jason Riesa, Dominika Rogozinska, Dror Marcus, Dalia El Badawy, Qiao Zhang, Luyu Wang, Helen Miller, Jeremy Greer, Lars Lowe Sjos, Azade Nova, Heiga Zen, Rahma Chaabouni, Mihaela Rosca, Jiepu Jiang, Charlie Chen, Ruibo Liu, Tara Sainath, Maxim Krikun, Alex Polozov, Jean-Baptiste Lespiau, Josh Newlan, Zeyncep Cankara, Soo Kwak, Yunhan Xu, Phil Chen, Andy Coenen, Clemens Meyer, Katerina Tsihlas, Ada Ma,

Juraj Gottweis, Jinwei Xing, Chenjie Gu, Jin Miao, Christian Frank, Zeynep Cankara, Sanjay Ganapathy, Ishita Dasgupta, Steph Hughes-Fitt, Heng Chen, David Reid, Keran Rong, Hongmin Fan, Joost van Amersfoort, Vincent Zhuang, Aaron Cohen, Shixiang Shane Gu, Anhad Mohananey, Anastasija Ilic, Taylor Tobin, John Wieting, Anna Bortsova, Phoebe Thacker, Emma Wang, Emily Caveness, Justin Chiu, Eren Sezener, Alex Kaskasoli, Steven Baker, Katie Millican, Mohamed Elhawaty, Kostas Aisopos, Carl Lebsack, Nathan Byrd, Hanjun Dai, Wenhao Jia, Matthew Wiethoff, Elnaz Davoodi, Albert Weston, Lakshman Yagati, Arun Ahuja, Isabel Gao, Golan Pundak, Susan Zhang, Michael Azzam, Khe Chai Sim, Sergi Caelles, James Keeling, Abhanshu Sharma, Andy Swing, YaGuang Li, Chenxi Liu, Carrie Grimes Bostock, Yamini Bansal, Zachary Nado, Ankesh Anand, Josh Lipschultz, Abhijit Karmarkar, Lev Proleev, Abe Ittycheriah, Soheil Hassas Yeganeh, George Polovets, Aleksandra Faust, Jiao Sun, Alban Rrustemi, Pen Li, Rakesh Shivanna, Jeremiah Liu, Chris Welty, Federico Lebron, Anirudh Baddepudi, Sebastian Krause, Emilio Parisotto, Radu Soricut, Zheng Xu, Dawn Bloxwich, Melvin Johnson, Behnam Neyshabur, Justin Mao-Jones, Renshen Wang, Vinay Ramasesh, Zaheer Abbas, Arthur Guez, Constant Segal, Duc Dung Nguyen, James Svensson, Le Hou, Sarah York, Kieran Milan, Sophie Bridgers, Wiktor Gworek, Marco Tagliasacchi, James Lee-Thorp, Michael Chang, Alexey Guseynov, Ale Jakse Hartman, Michael Kwong, Ruizhe Zhao, Sheleem Kashem, Elizabeth Cole, Antoine Miech, Richard Tanburn, Mary Phuong, Filip Pavetic, Sebastien Cevey, Ramona Comanescu, Richard Ives, Sherry Yang, Cosmo Du, Bo Li, Zizhao Zhang, Mariko Iinuma, Clara Huiyi Hu, Aurko Roy, Shaan Bijwadia, Zhenkai Zhu, Danilo Martins, Rachel Saputro, Anita Gergely, Steven Zheng, Dawei Jia, Ioannis Antonoglou, Adam Sadovsky, Shane Gu, Yingying Bi, Alek Andreev, Sina Samangooei, Mina Khan, Tomas Kocisky, Angelos Filos, Chintu Kumar, Colton Bishop, Adams Yu, Sarah Hodkinson, Sid Mittal, Premal Shah, Alexandre Moufarek, Yong Cheng, Adam Bloniarz, Jaehoon Lee, Pedram Pejman, Paul Michel, Stephen Spencer, Vladimir Feinberg, Xuehan Xiong, Nikolay Savinov, Charlotte Smith, Siamak Shakeri, Dustin Tran, Mary Chesus, Bernd Bohnet, George Tucker, Tamara von Glehn, Carrie Muir, Yiran Mao, Hideto Kazawa, Ambrose Slone, Kedar Soparkar, Disha Shrivastava, James Cobon-Kerr, Michael Sharman, Jay Pavagadhi, Carlos Araya, Karolis Misiunas, Nimesh Ghelani, Michael Laskin, David Barker, Qiujia Li, Anton Briukhov, Neil Houlsby, Mia Glaese, Balaji Lakshminarayanan, Nathan Schucher, Yunhao Tang, Eli Collins, Hyeontaek Lim, Fangxiaoyu Feng, Adria Recasens, Guangda Lai, Alberto Magni, Nicola De Cao, Aditya Siddhant, Zoe Ashwood, Jordi Orbay, Mostafa Dehghani, Jenny Brennan, Yifan He, Kelvin Xu, Yang Gao, Carl Saroufim, James Molloy, Xinyi Wu, Seb Arnold, Solomon Chang, Julian Schrittwieser, Elena Buchatskaya, Soroush Radpour, Martin Polacek, Skye Giordano, Ankur Bapna, Simon Tokumine, Vincent Hellendoorn, Thibault Sottiaux, Sarah Cogan, Aliaksei Severyn, Mohammad Saleh, Shantanu Thakoor, Laurent Shefey, Siyuan Qiao, Meenu Gaba, Shuo-yiin Chang, Craig Swanson, Biao Zhang, Benjamin Lee, Paul Kishan Rubenstein, Gan Song, Tom Kwiatkowski, Anna Koop, Ajay Kannan, David Kao, Parker Schuh, Axel Stjerngren, Golnaz Ghiasi, Gena Gibson, Luke Vilnis, Ye Yuan, Felipe Tiengo Ferreira, Aishwarya Kamath, Ted Klimenko, Ken Franko, Kefan Xiao, Indro Bhattacharya, Miteyan Patel, Rui Wang, Alex Morris, Robin Strudel, Vivek Sharma, Peter Choy, Sayed Hadi Hashemi, Jessica Landon, Mara Finkelstein, Priya Jhakra, Justin Frye, Megan Barnes, Matthew Mauger, Dennis Daun, Khuslen Baatarsukh, Matthew Tung, Wael Farhan, Henryk Michalewski, Fabio Viola, Felix de Chaumont Quitry, Charline Le Lan, Tom Hudson, Qingze Wang, Felix Fischer, Ivy Zheng, Elspeth White, Anca Dragan, Jean-baptiste Alayrac, Eric Ni, Alexander Pritzel, Adam Iwanicki, Michael Isard, Anna Bulanova, Lukas Zilka, Ethan Dyer, Devendra Sachan, Srivatsan Srinivasan, Hannah Muckenhirn, Honglong Cai, Amol Mandhane, Mukarram Tariq, Jack W. Rae, Gary Wang, Kareem Ayoub, Nicholas FitzGerald, Yao Zhao, Woohyun Han, Chris Alberti, Dan Garrette, Kashyap Krishnakumar, Mai Gimenez, Anselm Levskaya, Daniel Sohn, Josip Matak, Inaki Iturrate, Michael B. Chang, Jackie Xiang, Yuan Cao, Nishant Ranka, Geoff Brown, Adrian Hutter, Vahab Mirrokni, Nanxin Chen, Kaisheng Yao, Zoltan Egyed, Francois Galilee, Tyler Liechty, Praveen Kallakuri, Evan Palmer, Sanjay Ghemawat, Jasmine Liu, David Tao, Chloe Thornton, Tim Green, Mimi Jasarevic, Sharon Lin, Victor Cotruta, Yi-Xuan Tan, Noah Fiedel, Hongkun Yu, Ed Chi, Alexander Neitz, Jens Heitkaemper, Anu Sinha, Denny Zhou, Yi Sun, Charbel Kaed, Brice Hulse, Swaroop Mishra, Maria Georgaki, Sneha Kudugunta, Clement Farabet, Izhak Shafran, Daniel Vlasic, Anton Tsitsulin, Rajagopal Ananthanarayanan, Alen Carin, Guolong Su, Pei Sun, Shashank V, Gabriel Carvajal, Josef Broder, Iulia Comsa, Alena Repina, William Wong, Warren Weilun Chen, Peter Hawkins, Egor Filonov, Lucia Loher, Christoph Hirnschall, Weiyi Wang, Jingchen Ye, Andrea Burns, Hardie Cate, Diana Gage Wright, Federico Piccinini, Lei Zhang, Chu-Cheng Lin, Ionel Gog, Yana Kulizhskaya, Ashwin Sreevatsa, Shuang Song, Luis C. Cobo, Anand Iyer, Chetan Tekur, Guillermo Garrido, Zhuyun Xiao, Rupert Kemp, Huaixiu Steven Zheng, Hui Li, Ananth Agarwal, Christel Ngani, Kati Goshvadi, Rebeca Santamaria-Fernandez, Wojciech Fica, Xinyun Chen, Chris Gorgolewski, Sean Sun, Roopal Garg, Xinyu Ye, S. M. Ali Eslami, Nan Hua, Jon Simon, Pratik Joshi, Yelin Kim, Ian Tenney, Sahitya Potluri, Lam Nguyen Thiet, Quan Yuan, Florian Luisier, Alexandra Chronopoulou, Salvatore Scellato, Praveen Srinivasan, Minmin Chen, Vinod Koverkathu, Valentin Dalibard, Yaming Xu, Brennan Saeta, Keith Anderson, Thibault Sellam, Nick Fernando, Fantine Huot, Junehyuk Jung, Mani Varadarajan, Michael Quinn, Amit Raul, Maigo Le, Ruslan Habalov, Jon Clark, Komal Jalan, Kalesha Bullard, Achintya Singhal, Thang Luong, Boyu Wang, Sujeevan Rajayogam, Julian Eisenschlos, Johnson Jia, Daniel Finchelstein, Alex Yakubovich, Daniel Balle, Michael Fink, Sameer Agarwal, Jing Li, Dj Dvijotham, Shalini Pal, Kai Kang, Jaclyn Konzelmann, Jennifer Beattie, Olivier Dousse, Diane Wu, Remi Crocker, Chen Elkind, Siddhartha Reddy Jonnalagadda, Jong Lee, Dan Holtmann-Rice, Krystal Kallarackal, Rosanne Liu, Denis Vnukov, Neera Vats, Luca Invernizzi, Mohsen Jafari, Huanjie Zhou, Lilly Taylor, Jennifer Prendki, Marcus Wu, Tom Eccles, Tianqi Liu, Kavya Kopparapu, Francoise Beaufays, Christof Angermueller, Andreea Marzoca, Shourya Sarcar, Hilal Dib, Jeff Stanway, Frank Perbet, Nejc Trdin, Rachel Sterneck, Andrey Khorlin, Dinghua Li, Xihui Wu, Sonam Goenka, David Madras, Sasha Goldshtein, Willi Gierke, Tong Zhou, Yaxin Liu, Yannie Liang, Anais White, Yunjie Li, Shreya Singh, Sanaz Bahargam, Mark Epstein, Sujoy Basu, Li Lao, Adnan Ozturel, Carl Crous, Alex Zhai, Han Lu, Zora Tung, Neeraj Gaur, Alanna Walton, Lucas Dixon, Ming Zhang, Amir Globerson, Grant Uy, Andrew Bolt, Olivia Wiles, Milad Nasr, Ilia Shumailov, Marco Selvi, Francesco Piccinno, Ricardo Aguilar, Sara McCarthy, Misha Khalman, Mrinal Shukla, Vlado Galic, John Carpenter, Kevin Villela, Haibin Zhang, Harry Richardson, James Martens, Matko Bosnjak, Shreyas Rammohan Belle, Jeff Seibert, Mahmoud Alnahlawi, Brian McWilliams, Sankalp Singh, Annie Louis, Wen Ding, Dan Popovici, Lenin Simicich, Laura Knight, Pulkit Mehta, Nishesh Gupta, Chongyang Shi, Saaber Fatehi, Jovana Mitrovic, Alex Grills, Joseph Pagadora, Tsendsuren Munkhdalai, Dessie Petrova, Danielle Eisenbud, Zhishuai Zhang, Damion Yates, Bhavishya Mittal, Nilesh Tripuraneni, Yannis Assael, Thomas Brovelli, Prateek Jain, Mihajlo Velimirovic, Canfer Akbulut, Jiaqi Mu, Wolfgang Macherey, Ravin Kumar, Jun Xu, Haroon Qureshi, Gheorghe Comanici, Jeremy Wiesner, Zhitao Gong, Anton Ruddock, Matthias Bauer, Nick Felt, Anirudh GP, Anurag Arnab, Dustin Zelle, Jonas Rothfuss, Bill Rosgen, Ashish Shenoy, Bryan Seybold, Xinjian Li, Jayaram Mudigonda, Goker Erdogan, Jiawei Xia, Jiri Simsa, Andrea Michi, Yi Yao, Christopher Yew, Steven Kan, Isaac Caswell, Carey Radebaugh, Andre Elisseeff, Pedro Valenzuela, Kay McKinney, Kim Paterson, Albert Cui, Eri Latorre-Chimoto, Solomon Kim, William Zeng, Ken Durden, Priya Ponnapalli, Tiberiu Sosea, Christopher A. Choquette-Choo, James Manyika, Brona Robenek, Harsha Vashisht, Sebastien Pereira, Hoi Lam, Marko Velic, Denese Owusu-Afriyie, Katherine Lee, Tolga Bolukbasi, Alicia Parrish, Shawn Lu, Jane Park, Balaji Venkatraman, Alice Talbert, Lambert Rosique, Yuchung Cheng, Andrei Sozanschi, Adam Paszke, Praveen Kumar, Jessica Austin, Lu Li, Khalid Salama, Bartek Perz, Wooyeol Kim, Nandita Dukkipati, Anthony Baryshnikov, Christos Kaplanis, XiangHai Sheng, Yuri Chervonyi, Caglar Unlu, Diego de Las Casas, Harry Askham, Kathryn Tunyasuvunakool, Felix Gimeno, Siim Poder, Chester Kwak, Matt Miecnikowski, Vahab Mirrokni, Alek Dimitriev, Aaron Parisi, Dangyi Liu, Tomy Tsai, Toby Shevlane, Christina Kouridi, Drew Garmon, Adrian Goedeckemeyer, Adam R. Brown, Anitha Vijayakumar, Ali Elqursh, Sadegh Jazayeri, Jin Huang, Sara Mc Carthy, Jay Hoover, Lucy Kim, Sandeep Kumar, Wei Chen, Courtney Biles, Garrett Bingham, Evan Rosen, Lisa Wang, Qijun Tan, David Engel, Francesco Pongetti, Dario de Cesare, Dongseong Hwang, Lily Yu, Jennifer Pullman, Srini Narayanan, Kyle Levin, Siddharth Gopal, Megan Li, Asaf Aharoni, Trieu Trinh, Jessica Lo, Norman Casagrande, Roopali Vij, Loic Matthey, Bramandia Ramadhana, Austin Matthews, C. J. Carey, Matthew Johnson, Kremena Goranova, Rohin Shah, Shereen Ashraf, Kingshuk Dasgupta, Rasmus Larsen, Yicheng Wang, Manish Reddy Vuyyuru, Chong Jiang, Joana Ijazi, Kazuki Osawa, Celine Smith, Ramya Sree Boppana, Taylan Bilal, Yuma Koizumi, Ying Xu, Yasemin Altun, Nir Shabat, Ben Bariach, Alex Korchemniy, Kiam Choo, Olaf Ronneberger, Chimezie Iwuanyanwu, Shubin Zhao, David Soergel, Cho-Jui Hsieh, Irene Cai, Shariq Iqbal, Martin Sundermeyer, Zhe Chen, Elie Bursztein, Chaitanya Malaviya, Fadi Biadsy, Prakash Shroff, Inderjit Dhillon, Tejasi Latkar, Chris Dyer, Hannah Forbes, Massimo Nicosia, Vitaly Nikolaev, Somer Greene, Marin Georgiev, Pidong Wang, Nina Martin, Hanie Sedghi, John Zhang, Praseem Banzal, Doug Fritz, Vikram Rao, Xuezhi Wang, Jiageng Zhang, Viorica Patraucean, Dayou Du, Igor Mordatch, Ivan Jurin, Lewis Liu, Ayush Dubey, Abhi Mohan, Janek Nowakowski, Vlad-Doru Ion, Nan Wei, Reiko Tojo, Maria Abi Raad, Drew A. Hudson, Vaishakh Keshava, Shubham Agrawal, Kevin Ramirez, Zhichun Wu, Hoang Nguyen, Ji Liu, Madhavi Sewak, Bryce Petrini, DongHyun Choi, Ivan Philips, Ziyue Wang, Ioana Bica, Ankush Garg, Jarek Wilkiewicz, Priyanka Agrawal, Xiaowei Li, Danhao Guo, Emily Xue, Naseer Shaik, Andrew Leach, Sadh MNM Khan, Julia Wiesinger, Sammy Jerome, Abhishek Chakladar, Alek Wenjiao Wang, Tina Ornduff, Folake Abu, Alireza Ghaffarkhah, Marcus Wainwright, Mario Cortes, Frederick Liu, Joshua Maynez, Andreas Terzis, Pouya Samangouei, Riham Mansour, Tomasz Kępa, François-Xavier Aubet, Anton Algymr, Dan Banica, Agoston Weisz, Andras Orban, Alexandre Senges, Ewa Andrejczuk, Mark Geller, Niccolo Dal Santo, Valentin Anklin, Majd Al Merey, Martin Baeuml, Trevor Strohman, Junwen Bai, Slav Petrov, Yonghui Wu, Demis Hassabis, Koray Kavukcuoglu, Jeff Dean, and Oriol Vinyals. 2024. Gemini 1.5: Unlocking Multimodal Understanding across Millions of Tokens of Context. https://doi.org/10.48550/arXiv.2403.05530 arXiv:2403.05530 [cs]

[205] Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, Pouya Tafti, Léonard Hussenot, Pier Giuseppe Sessa, Aakanksha Chowdhery, Adam Roberts, Aditya Barua, Alex Botev, Alex Castro-Ros, Ambrose Slone, Amélie Héliou, Andrea Tacchetti, Anna Bulanova, Antonia Paterson, Beth Tsai, Bobak Shahriari, Charline Le Lan, Christopher A. Choquette-Choo, Clément Crepy, Daniel Cer, Daphne Ippolito, David Reid, Elena Buchatskaya, Eric Ni, Eric Noland, Geng Yan, George Tucker, George-Christian Muraru, Grigory Rozhdestvenskiy, Henryk Michalewski, Ian Tenney, Ivan Grishchenko, Jacob Austin, James Keeling, Jane Labanowski, Jean-Baptiste Lespiau, Jeff Stanway, Jenny Brennan, Jeremy Chen, Johan Ferret, Justin Chiu, Justin Mao-Jones, Katherine Lee, Kathy Yu, Katie Millican, Lars Lowe Sjoesund, Lisa Lee, Lucas Dixon, Machel Reid, Maciej Mikuła, Mateo Wirth, Michael Sharman, Nikolai Chinaev, Nithum Thain, Olivier Bachem, Oscar Chang, Oscar Wahltinez, Paige Bailey, Paul Michel, Petko Yotov, Rahma Chaabouni, Ramona Comanescu, Reena Jana, Rohan Anil, Ross McIlroy, Ruibo Liu, Ryan Mullins, Samuel L. Smith, Sebastian Borgeaud, Sertan Girgin, Sholto Douglas, Shree Pandya, Siamak Shakeri, Soham De, Ted Klimenko, Tom Hennigan, Vlad Feinberg, Wojciech Stokowiec, Yu-hui Chen, Zafarali Ahmed, Zhitao Gong, Tris Warkentin, Ludovic Peran, Minh Giang, Clément Farabet, Oriol Vinyals, Jeff Dean, Koray Kavukcuoglu, Demis Hassabis, Zoubin Ghahramani, Douglas Eck, Joelle Barral, Fernando Pereira, Eli Collins, Armand Joulin, Noah Fiedel, Evan Senter, Alek Andreev, and Kathleen Kenealy. 2024. Gemma: Open Models Based on Gemini Research and Technology. https://doi.org/10.48550/arXiv.2403.08295 arXiv:2403.08295 [cs]

[206] Divy Thakkar, Azra Ismail, Pratyush Kumar, Alex Hanna, Nithya Sambasivan, and Neha Kumar. 2022. When Is Machine Learning Data Good?: Valuing in Public Health Datafication. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems (CHI '22)*. Association for Computing Machinery, New York, NY, USA, 1–16. https://doi.org/10.1145/3491102.3501868

[207] Eva Thelisson, Kshitij Sharma, Hanan Salam, and Virginia Dignum. 2018. The General Data Protection Regulation: An Opportunity for the HCI Community?. In *Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems (CHI EA '18)*. Association for Computing Machinery, New York, NY, USA, 1–8. https://doi.org/10.1145/3170427.3170632

[208] Tim Theys, Stephanie Van Hove, Peter Mechant, Gill Van Impe, Alexander Heerinckx, and Jelle Saldien. 2025. Exploring Users' Perspectives on a Solid-Enabled Personal Data Store Enhanced Streaming Service. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI '25)*. Association for Computing Machinery, New York, NY, USA, 1–22. https://doi.org/10.1145/3706598.3713308

[209] David Thiel. 2023. Investigation Finds AI Image Generation Models Trained on Child Abuse.

[210] Lauren Thornton, Bran Knowles, and Gordon Blair. 2021. Fifty Shades of Grey: In Praise of a Nuanced Approach Towards Trustworthy Design. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT '21)*. Association for Computing Machinery, New York, NY, USA, 64–76. https://doi.org/10.1145/3442188.3445871

[211] Bill Tomlinson, Donald J. Patterson, and Andrew W. Torrance. 2023. Turning Fake Data into Fake News: The AI Training Set as a Trojan Horse of Misinformation. *San Diego Law Review* 60 (2023), 641.

[212] Tolgahan Toy. 2023. Transparency in AI. *AI & SOCIETY* (Oct. 2023). https://doi.org/10.1007/s00146-023-01786-y

[213] Emily Tseng, Meg Young, Marianne Aubin Le Quéré, Aimee Rinehart, and Harini Suresh. 2025. "Ownership, Not Just Happy Talk": Co-Designing a Participatory Large Language Model for Journalism. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency (FAccT '25)*. Association for Computing Machinery, New York, NY, USA, 3119–3130. https://doi.org/10.1145/3715275.3732198

[214] Matteo Turilli and Luciano Floridi. 2009. The Ethics of Information Transparency. *Ethics and Information Technology* 11, 2 (June 2009), 105–112. https://doi.org/10.1007/s10676-009-9187-9

[215] Michel E. van Genderen, Davy van de Sande, Lotty Hooft, Andreas Alois Reis, Alexander D. Cornet, Jacobien H. F. Oosterhoff, Björn J. P. van der Ster, Joost Huiskens, Reggie Townsend, Jasper van Bommel, Diederik Gommers, and Jeroen van den Hoven. 2024. Charting a New Course in Healthcare: Early-Stage AI Algorithm Registration to Enhance Trust and Transparency. *npj Digital Medicine* 7, 1 (May 2024), 1–4. https://doi.org/10.1038/s41746-024-01104-w

[216] Jennifer Wortman Vaughan and Hanna Wallach. 2021. A Human-Centered Agenda for Intelligible Machine Learning. In *In Machines We Trust: Perspectives on Dependable AI*. MIT Press.

[217] Michael Veale, Max Van Kleek, and Reuben Binns. 2018. Fairness and Accountability Design Needs for Algorithmic Support in High-Stakes Public Sector Decision-Making. *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (2018). https://doi.org/10.1145/3173574

[218] Pieter Verdegem. 2024. Dismantling AI Capitalism: The Commons as an Alternative to the Power Concentration of Big Tech. *AI & SOCIETY* 39, 2 (April 2024), 727–737. https://doi.org/10.1007/s00146-022-01437-8

[219] Pablo Villalobos, Anson Ho, Jaime Sevilla, Tamay Besiroglu, Lennart Heim, and Marius Hobbhahn. 2024. Position: Will We Run out of Data? Limits of LLM Scaling Based on Human-Generated Data. In *Forty-First International Conference on Machine Learning*.

[220] Warren J. von Eschenbach. 2021. Transparency and the Black Box Problem: Why We Do Not Trust AI. *Philosophy & Technology* 34, 4 (Dec. 2021), 1607–1622. https://doi.org/10.1007/s13347-021-00477-0

[221] Krzysztof Wach, Cong Doanh Duong, Joanna Ejdys, Rūta Kazlauskaitė, Pawel Korzynski, Grzegorz Mazurek, Joanna Paliszkiewicz, and Ewa Ziemba. 2023. The Dark Side of Generative Artificial Intelligence: A Critical Analysis of Controversies and Risks of ChatGPT. *Entrepreneurial Business and Economics Review* 11, 2 (June 2023), 7–30. https://doi.org/10.15678/EBER.2023.110201

[222] Angelina Wang. 2025. Identities Are Not Interchangeable: The Problem of Overgeneralization in Fair Machine Learning. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency (FAccT '25)*. Association for Computing Machinery, New York, NY, USA, 485–497. https://doi.org/10.1145/3715275.3732033

[223] Weilin Wang. 2025. Research on Copyright of Generative Artificial Intelligence for Promoting Data Open Utilization. *Modern Law Research* 6, 1 (June 2025). https://doi.org/10.37420/j.mlr.2025.013

[224] Yisen Wang. 2024. Secure and Trustworthy Large Language Models. In *ICLR 2024 Workshops*.

[225] Stephen J. A. Ward. 2014. The Magical Concept of Transparency. In *Ethics for Digital Journalists*. Routledge.

[226] Maurice Weber, Daniel Fu, Quentin Anthony, Yonatan Oren, Shane Adams, Anton Alexandrov, Xiaozhong Lyu, Huu Nguyen, Xiaozhe Yao, Virginia Adams, Ben Athiwaratkun, Rahul Chalamala, Kezhen Chen, Max Ryabinin, Tri Dao, Percy Liang, Christopher Ré, Irina Rish, and Ce Zhang. 2024. RedPajama: An Open Dataset for Training Large Language Models. https://doi.org/10.48550/arXiv.2411.12372 arXiv:2411.12372 [cs]

[227] David Weil. 2002. The Benefits and Costs of Transparency: A Model of Disclosure Based Regulation. https://doi.org/10.2139/ssrn.316145 social science research network:316145

[228] Jack West, Bengisu Cagiltay, Shirley Zhang, Jingjie Li, Kassem Fawaz, and Suman Banerjee. 2025. "Impressively Scary:' Exploring User Perceptions and Reactions to Unraveling Machine Learning Models in Social Media Applications. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI '25)*. Association for Computing Machinery, New York, NY, USA, 1–21. https://doi.org/10.1145/3706598.3713256

[229] Matt White, Ibrahim Haddad, Cailean Osborne, Xiao-Yang Yanglet Liu, Ahmed Abdelmonsef, Sachin Varghese, and Arnaud Le Hors. 2024. The Model Openness Framework: Promoting Completeness and Openness for Reproducibility, Transparency, and Usability in Artificial Intelligence. https://doi.org/10.48550/arXiv.2403.13784 arXiv:2403.13784

[230] David Gray Widder, Dawn Nafus, Laura Dabbish, and James Herbsleb. 2022. Limits and Possibilities for "Ethical AI" in Open Source: A Study of Deepfakes. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT '22)*. Association for Computing Machinery, New York, NY, USA, 2035–2046. https://doi.org/10.1145/3531146.3533779

[231] David Gray Widder, Meredith Whittaker, and Sarah Myers West. 2024. Why 'Open' AI Systems Are Actually Closed, and Why This Matters. *Nature* 635, 8040 (Nov. 2024), 827–833. https://doi.org/10.1038/s41586-024-08141-1

[232] Mark D. Wilkinson, Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, Jan-Willem Boiten, Luiz Bonino da Silva Santos, Philip E. Bourne, Jildau Bouwman, Anthony J. Brookes, Tim Clark, Mercè Crosas, Ingrid Dillo, Olivier Dumon, Scott Edmunds, Chris T. Evelo, Richard Finkers, Alejandra Gonzalez-Beltran, Alasdair J. G. Gray, Paul Groth, Carole Goble, Jeffrey S. Grethe, Jaap Heringa, Peter A. C. 't Hoen, Rob Hooft, Tobias Kuhn, Ruben Kok, Joost Kok, Scott J. Lusher, Maryann E. Martone, Albert Mons, Abel L. Packer, Bengt Persson, Philippe Rocca-Serra, Marco Roos, Rene van Schaik, Susanna-Assunta Sansone, Erik Schultes, Thierry Sengstag, Ted Slater, George Strawn, Morris A. Swertz, Mark Thompson, Johan van der Lei, Erik van Mulligen, Jan Velterop, Andra Waagmeester, Peter Wittenburg, Katherine Wolstencroft, Jun Zhao, and Barend Mons. 2016. The FAIR Guiding Principles for Scientific Data Management and Stewardship. *Scientific Data* 3, 1 (March 2016), 160018. https://doi.org/10.1038/sdata.2016.18

[233] Amy Winecoff and Miranda Bogen. 2025. Improving Governance Outcomes Through AI Documentation: Bridging Theory and Practice. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI '25)*. Association for Computing Machinery, New York, NY, USA, 1–18. https://doi.org/10.1145/3706598.3713814

[234] BigScience Workshop, Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, Jonathan Tow, Alexander M. Rush, Stella Biderman, Albert Webson, Pawan Sasanka Ammanamanchi, Thomas Wang, Benoît Sagot, Niklas Muennighoff, Albert Villanova del Moral, Olatunji Ruwase, Rachel Bawden, Stas Bekman, Angelina McMillan-Major, Iz Beltagy, Huu Nguyen, Lucile Saulnier, Samson Tan, Pedro Ortiz Suarez, Victor Sanh, Hugo Laurençon,

Yacine Jernite, Julien Launay, Margaret Mitchell, Colin Raffel, Aaron Gokaslan, Adi Simhi, Aitor Soroa, Alham Fikri Aji, Amit Alfassy, Anna Rogers, Ariel Kreisberg Nitzav, Canwen Xu, Chenghao Mou, Chris Emezue, Christopher Klamm, Colin Leong, Daniel van Strien, David Ifeoluwa Adelani, Dragomir Radev, Eduardo González Ponferrada, Efrat Levkovizh, Ethan Kim, Eyal Bar Natan, Francesco De Toni, Gérard Dupont, Germán Kruszewski, Giada Pistilli, Hady Elsahar, Hamza Benyamina, Hieu Tran, Ian Yu, Idris Abdulmumin, Isaac Johnson, Itziar Gonzalez-Dios, Javier de la Rosa, Jenny Chim, Jesse Dodge, Jian Zhu, Jonathan Chang, Jörg Frohberg, Joseph Tobing, Joydeep Bhattacharjee, Khalid Almubarak, Kimbo Chen, Kyle Lo, Leandro Von Werra, Leon Weber, Long Phan, Loubna Ben allal, Ludovic Tanguy, Manan Dey, Manuel Romero Muñoz, Maraim Masoud, María Grandury, Mario Šaško, Max Huang, Maximin Coavoux, Mayank Singh, Mike Tian-Jian Jiang, Minh Chien Vu, Mohammad A. Jauhar, Mustafa Ghaleb, Nishant Subramani, Nora Kassner, Nurulaqilla Khamis, Olivier Nguyen, Omar Espejel, Ona de Gibert, Paulo Villegas, Peter Henderson, Pierre Colombo, Priscilla Amuok, Quentin Lhoest, Rheza Harliman, Rishi Bommasani, Roberto Luis López, Rui Ribeiro, Salomey Osei, Sampo Pyysalo, Sebastian Nagel, Shamik Bose, Shamsuddeen Hassan Muhammad, Shanya Sharma, Shayne Longpre, Somaieh Nikpoor, Stanislav Silberberg, Suhas Pai, Sydney Zink, Tiago Timponi Torrent, Timo Schick, Tristan Thrush, Valentin Danchev, Vassilina Nikoulina, Veronika Laippala, Violette Lepercq, Vrinda Prabhu, Zaid Alyafeai, Zeerak Talat, Arun Raja, Benjamin Heinzerling, Chenglei Si, Davut Emre Taşar, Elizabeth Salesky, Sabrina J. Mielke, Wilson Y. Lee, Abheesht Sharma, Andrea Santilli, Antoine Chaffin, Arnaud Stiegler, Debajyoti Datta, Eliza Szczechla, Gunjan Chhablani, Han Wang, Harshit Pandey, Hendrik Strobelt, Jason Alan Fries, Jos Rozen, Leo Gao, Lintang Sutawika, M. Saiful Bari, Maged S. Al-shaibani, Matteo Manica, Nihal Nayak, Ryan Teehan, Samuel Albanie, Sheng Shen, Srulik Ben-David, Stephen H. Bach, Taewoon Kim, Tali Bers, Thibault Fevry, Trishala Neeraj, Urmish Thakker, Vikas Raunak, Xiangru Tang, Zheng-Xin Yong, Zhiqing Sun, Shaked Brody, Yallow Uri, Hadar Tojarieh, Adam Roberts, Hyung Won Chung, Jaesung Tae, Jason Phang, Ofir Press, Conglong Li, Deepak Narayanan, Hatim Bourfoune, Jared Casper, Jeff Rasley, Max Ryabinin, Mayank Mishra, Minjia Zhang, Mohammad Shoeybi, Myriam Peyrounette, Nicolas Patry, Nouamane Tazi, Omar Sanseviero, Patrick von Platen, Pierre Cornette, Pierre François Lavallée, Rémi Lacroix, Samyam Rajbhandari, Sanchit Gandhi, Shaden Smith, Stéphane Requena, Suraj Patil, Tim Dettmers, Ahmed Baruwa, Amanpreet Singh, Anastasia Cheveleva, Anne-Laure Ligozat, Arjun Subramonian, Aurélie Névéol, Charles Lovering, Dan Garrette, Deepak Tunuguntla, Ehud Reiter, Ekaterina Taktasheva, Ekaterina Voloshina, Eli Bogdanov, Genta Indra Winata, Hailey Schoelkopf, Jan-Christoph Kalo, Jekaterina Novikova, Jessica Zosa Forde, Jordan Clive, Jungo Kasai, Ken Kawamura, Liam Hazan, Marine Carpuat, Miruna Clinciu, Najoung Kim, Newton Cheng, Oleg Serikov, Omer Antverg, Oskar van der Wal, Rui Zhang, Ruochen Zhang, Sebastian Gehrmann, Shachar Mirkin, Shani Pais, Tatiana Shavrina, Thomas Scialom, Tian Yun, Tomasz Limisiewicz, Verena Rieser, Vitaly Protasov, Vladislav Mikhailov, Yada Pruksachatkun, Yonatan Belinkov, Zachary Bamberger, Zdeněk Kasner, Alice Rueda, Amanda Pestana, Amir Feizpour, Ammar Khan, Amy Faranak, Ana Santos, Anthony Hevia, Antigona Unldreaj, Arash Aghagol, Arezoo Abdollahi, Aycha Tammour, Azadeh HajiHosseini, Bahareh Behroozi, Benjamin Ajibade, Bharat Saxena, Carlos Muñoz Ferrandis, Daniel McDuff, Danish Contractor, David Lansky, Davis David, Douwe Kiela, Duong A. Nguyen, Edward Tan, Emi Baylor, Ezinwanne Ozoani, Fatima Mirza, Frankline Ononiwu, Habib Rezanejad, Hessie Jones, Indrani Bhattacharya, Irene Solaiman, Irina Sedenko, Isar Nejadgholi, Jesse Passmore, Josh Seltzer, Julio Bonis Sanz, Livia Dutra, Mairon Samagaio, Maraim Elbadri, Margot Mieskes, Marissa Gerchick, Martha Akinlolu, Michael McKenna, Mike Qiu, Muhammed Ghauri, Mykola Burynok, Nafis Abrar, Nazneen Rajani, Nour Elkott, Nour Fahmy, Olanrewaju Samuel, Ran An, Rasmus Kromann, Ryan Hao, Samira Alizadeh, Sarmad Shubber, Silas Wang, Sourav Roy, Sylvain Viguier, Thanh Le, Tobi Oyebade, Trieu Le, Yoyo Yang, Zach Nguyen, Abhinav Ramesh Kashyap, Alfredo Palasciano, Alison Callahan, Anima Shukla, Antonio Miranda-Escalada, Ayush Singh, Benjamin Beilharz, Bo Wang, Caio Brito, Chenxi Zhou, Chirag Jain, Chuxin Xu, Clémentine Fourrier, Daniel León Periñán, Daniel Molano, Dian Yu, Enrique Manjavacas, Fabio Barth, Florian Fuhrimann, Gabriel Altay, Giyaseddin Bayrak, Gully Burns, Helena U. Vrabec, Imane Bello, Ishani Dash, Jihyun Kang, John Giorgi, Jonas Golde, Jose David Posada, Karthik Rangasai Sivaraman, Lokesh Bulchandani, Lu Liu, Luisa Shinzato, Madeleine Hahn de Bykhovetz, Maiko Takeuchi, Marc Pàmies, Maria A. Castillo, Marianna Nezhurina, Mario Sänger, Matthias Samwald, Michael Cullan, Michael Weinberg, Michiel De Wolf, Mina Mihaljcic, Minna Liu, Moritz Freidank, Myungsun Kang, Natasha Seelam, Nathan Dahlberg, Nicholas Michio Broad, Nikolaus Muellner, Pascale Fung, Patrick Haller, Ramya Chandrasekhar, Renata Eisenberg, Robert Martin, Rodrigo Canalli, Rosaline Su, Ruisi Su, Samuel Cahyawijaya, Samuele Garda, Shlok S. Deshmukh, Shubhanshu Mishra, Sid Kiblawi, Simon Ott, Sinee Sang-aroonsiri, Srishti Kumar, Stefan Schweter, Sushil Bharati, Tanmay Laud, Théo Gigant, Tomoya Kainuma, Wojciech Kusa, Yanis Labrak, Yash Shailesh Bajaj, Yash Venkatraman, Yifan Xu, Yingxin Xu, Yu Xu, Zhe Tan, Zhongli Xie, Zifan Ye, Mathilde Bras, Younes Belkada, and Thomas Wolf. 2023. BLOOM: A 176B-Parameter Open-Access Multilingual Language Model. https://doi.org/10.48550/arXiv.2211.05100 arXiv:2211.05100 [cs]

[235] Sophia Worth, Ben Snaith, Arunav Das, Gefion Thuermer, and Elena Simperl. 2024. AI Data Transparency: An Exploration through the Lens of AI Incidents. https://doi.org/10.48550/arXiv.2409.03307 arXiv:2409.03307

[236] Chao Wu. 2024. Data Privacy: From Transparency to Fairness. *Technology in Society* 76 (March 2024), 102457. https://doi.org/10.1016/j.techsoc.2024.102457

[237] Mengke Wu, Weizi Liu, Yanyun Wang, and Mike Yao. 2025. Negotiating the Shared Agency between Humans & AI in the Recommender System. In *Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems (CHI EA '25)*. Association for Computing Machinery, New York, NY, USA, 1–9. https://doi.org/10.1145/3706599.3719900

[238] Yuxin Xiao, Shulammite Lim, Tom Joseph Pollard, and Marzyeh Ghassemi. 2023. In the Name of Fairness: Assessing the Bias in Clinical Record De-identification. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency (FAccT '23)*. Association for Computing Machinery, New York, NY, USA, 123–137. https://doi.org/10.1145/3593013.3593982

[239] Zhihan Xu and Eni Mustafaraj. 2024. Tracing the Evolution of Information Transparency for OpenAI's GPT Models through a Biographical Approach. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society* 7 (Oct. 2024), 1684–1695. https://doi.org/10.1609/aies.v7i1.31757

[240] Xi Yang and Marco Aurisicchio. 2021. Designing Conversational Agents: A Self-Determination Theory Approach. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems (CHI '21)*. Association for Computing Machinery, New York, NY, USA, 1–16. https://doi.org/10.1145/3411764.3445445

[241] Dong Whi Yoo, Hayoung Woo, Viet Cuong Nguyen, Michael L. Birnbaum, Kaylee Payne Kruzan, Jennifer G Kim, Gregory D. Abowd, and Munmun De Choudhury. 2024. Patient Perspectives on AI-Driven Predictions of Schizophrenia Relapses: Understanding Concerns and Opportunities for Self-Care and Treatment. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems (CHI '24)*. Association for Computing Machinery, New York, NY, USA, 1–20. https://doi.org/10.1145/3613904.3642369

[242] Meg Young, Luke Rodriguez, Emily Keller, Feiyang Sun, Boyang Sa, Jan Whittington, and Bill Howe. 2019. Beyond Open vs. Closed: Balancing Individual Privacy and Public Accountability in Data Sharing. In *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT* '19)*. Association for Computing Machinery, New York, NY, USA, 191–200. https://doi.org/10.1145/3287560.3287577

[243] Xiao Yu, Zexian Zhang, Feifei Niu, Xing Hu, Xin Xia, and John Grundy. 2024. What Makes a High-Quality Training Dataset for Large Language Models: A Practitioners' Perspective. In *Proceedings of the 39th IEEE/ACM International Conference on Automated Software Engineering (ASE '24)*. Association for Computing Machinery, New York, NY, USA, 656–668. https://doi.org/10.1145/3691620.3695061

[244] Monika Zalnieriute. 2021. "Transparency Washing" in the Digital Age: A Corporate Agenda of Procedural Fetishism. *Critical Analysis of Law* 8, 1 (April 2021), 139–153. https://doi.org/10.33137/cal.v8i1.36284

[245] He Zhang, Siyu Zha, Jie Cai, Donghee Yvette Wohn, and John M. Carroll. 2025. Generative AI in Virtual Reality Communities: A Preliminary Analysis of the VRChat Discord Community. In *Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems (CHI EA '25)*. Association for Computing Machinery, New York, NY, USA, 1–11. https://doi.org/10.1145/3706599.3720120

[246] Banghua Zhu, Evan Frick, Tianhao Wu, Hanlin Zhu, Karthik Ganesan, Wei-Lin Chiang, Jian Zhang, and Jiantao Jiao. 2024. Starling-7B: Improving Helpfulness and Harmlessness with RLAIF. In *First Conference on Language Modeling*.

[247] Haiyi Zhu, Bowen Yu, Aaron Halfaker, and Loren Terveen. 2018. Value-Sensitive Algorithm Design: Method, Case Study, and Lessons. *Proceedings of the ACM on Human-Computer Interaction* 2, CSCW (2018). https://doi.org/10.1145/3274463

[248] Maciej Krzysztof Zuziak, Onntje Hinrichs, Aizhan Abdrassulova, and Salvatore Rinzivillo. 2023. Data Collaboratives with the Use of Decentralised Learning. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency (FAccT '23)*. Association for Computing Machinery, New York, NY, USA, 615–625. https://doi.org/10.1145/3593013.3594029

# A Appendix

## A.1 Document Selection Details

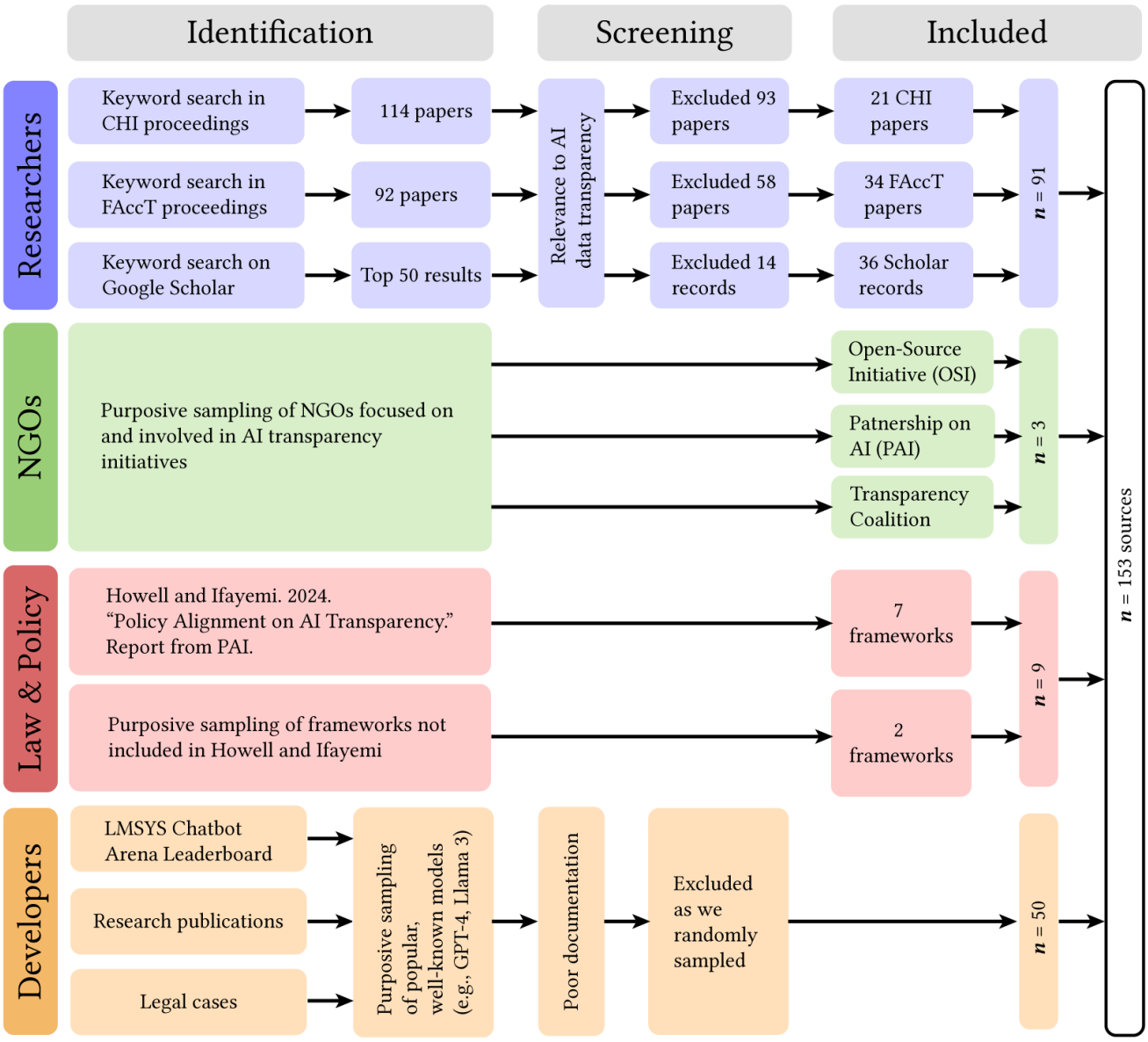



**Figure 1: An altered PRISMA Flow Diagram [156] showing the sampling processes we took for each of the four stakeholder groups we sourced documentation from. Given documentation is shared much different between these four types of stakeholders, we took different approaches to sampling the appropriate representative sources.**

### *A.1.1 PRISMA Flow Diagram.*

### *A.1.2 Documentation Corpus List.*

| *Stakeholder Role* | *Source* | *Citation / Name* |
|---|---|---|
| Researchers | CHI | [27, 29, 62, 100, 107, 114, 126, 141, 187, 188, 193, 206–208, 217, 228, 233, 237, 240, 241, 245] |
| | FAccT | [19, 34, 36, 40, 57, 58, 64, 65, 68, 72, 77, 81, 94, 103, 106, 111, 120, 128, 150, 158, 160, 164, 165, 167, 175, 181, 194, 210, 213, 230, 238, 242, 248] |
| | Scholar | [21, 28, 46–49, 70, 74, 76, 86, 87, 92, 95, 98, 101, 105, 118, 129–131, 133, 143, 145, 151, 163, 174, 184, 190, 191, 203, 212, 215, 229, 235, 236, 239] |
| NGOs | Purposive sampling | Open-Source Initiative (OSI), Parnership on AI (PAI), The Transparency Coalition |
| Law & Policy Leaders | [104] | [4, 8, 12–14, 154, 154] |
| | Purposive sampling | [5, 9] |
| Developers | Various sources, including: LMSYS Chatbot Arena Leaderboard & [33, 47, 53, 197, 198] | Alpaca, Amazon Nova, Arctic-Embed, Aya, BLOOM, Claude 3, Command A, DBRX, Deepseek-V3, Dolphin, ERNIE 3.0 Titan, Falcon, Gemini, Gemini 1.5, Gemma, Gemma 2, GLM, GPT-1, GPT-2, GPT-3, GPT-4, Granite, Guanaco, Hermes 3, HunyuanVideo, InternLM2, Jamba, KOALA, LLaMA, Llama 2, Llama 3, Mistral 7B, Mixtral 8x7B, Nemotron-4, OLMo, OpenChat, PaLM 2, Phi-3, Qwen3, RedPajama-INCITE, Reka, Sparrow, Stable Diffusion 3, Stable LM 2, Starling 7B, Step-Audio, Tulu 3, Vicuna, Yi, Zephyr |

**Table 4: A table documenting each of our main sources in our document analysis for each stakeholder type, as described in Section 3.2.**

## A.2 Query Syntaxes

*A.2.1 FAccT ACM Full Query Syntax.* "query": Title:(dataset transparency) OR Title:(data sharing) OR Title:(open data) OR Abstract:(dataset transparency) OR Abstract:(data sharing) OR Abstract:(open data) OR Keyword:(dataset transparency) OR Keyword:(data sharing) OR Keyword:(open data) "filter": Conference Collections: FAccT: Fairness, Accountability, and Transparency

*A.2.2 CHI ACM Full Query Syntax.* "query": Title:(data transparency) OR Abstract:(data transparency) OR Keyword:(data transparency) "filter": Conference Collections: CHI: Conference on Human Factors in Computing Systems

## A.3 Additional Figures

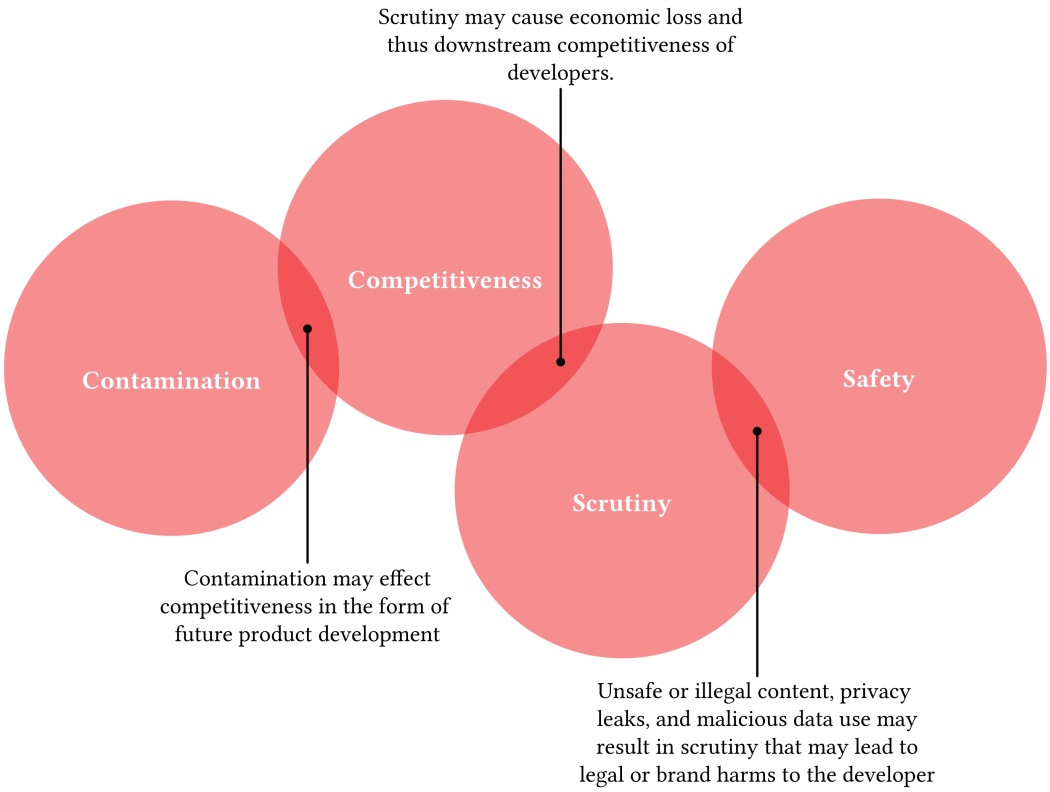



**Figure 2: The figure shows how the risks we identified can overlap with one another. We note that these overlaps may not be exhaustive; other overlapping considerations between each risk or multiple risks may also exist.**

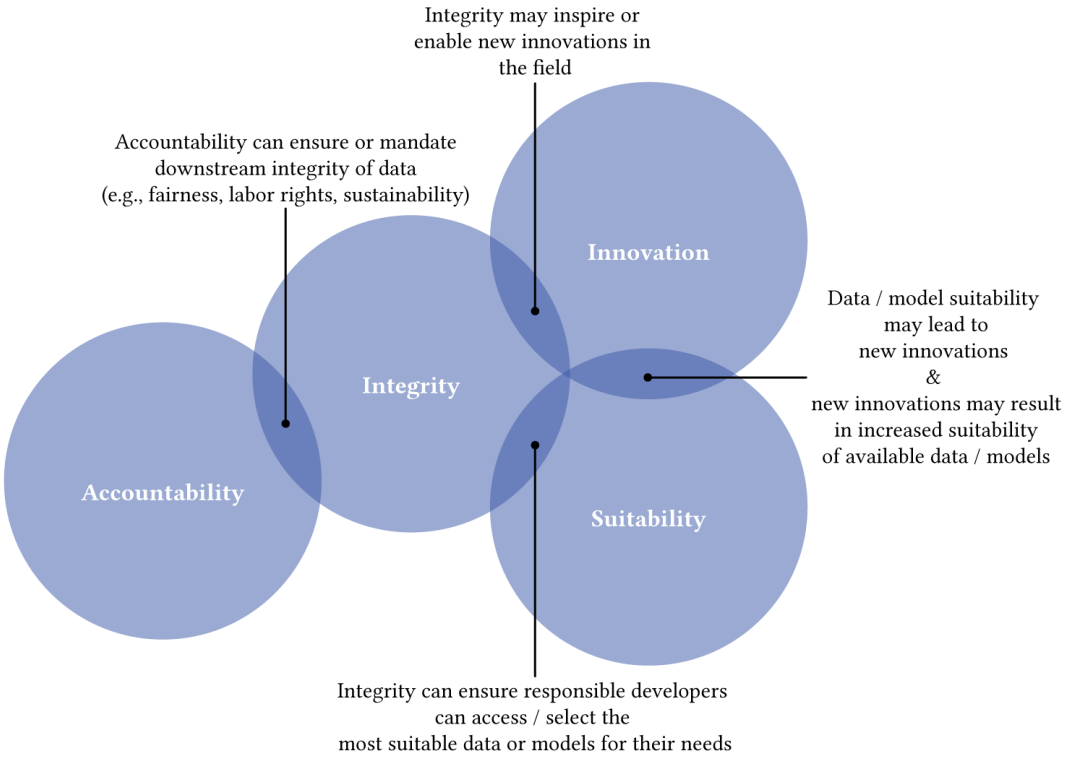



**Figure 3: The figure shows how the benefits we identified can overlap with one another. We note that these overlaps may not be exhaustive; other overlapping considerations between each benefit or multiple benefits may also exist.**

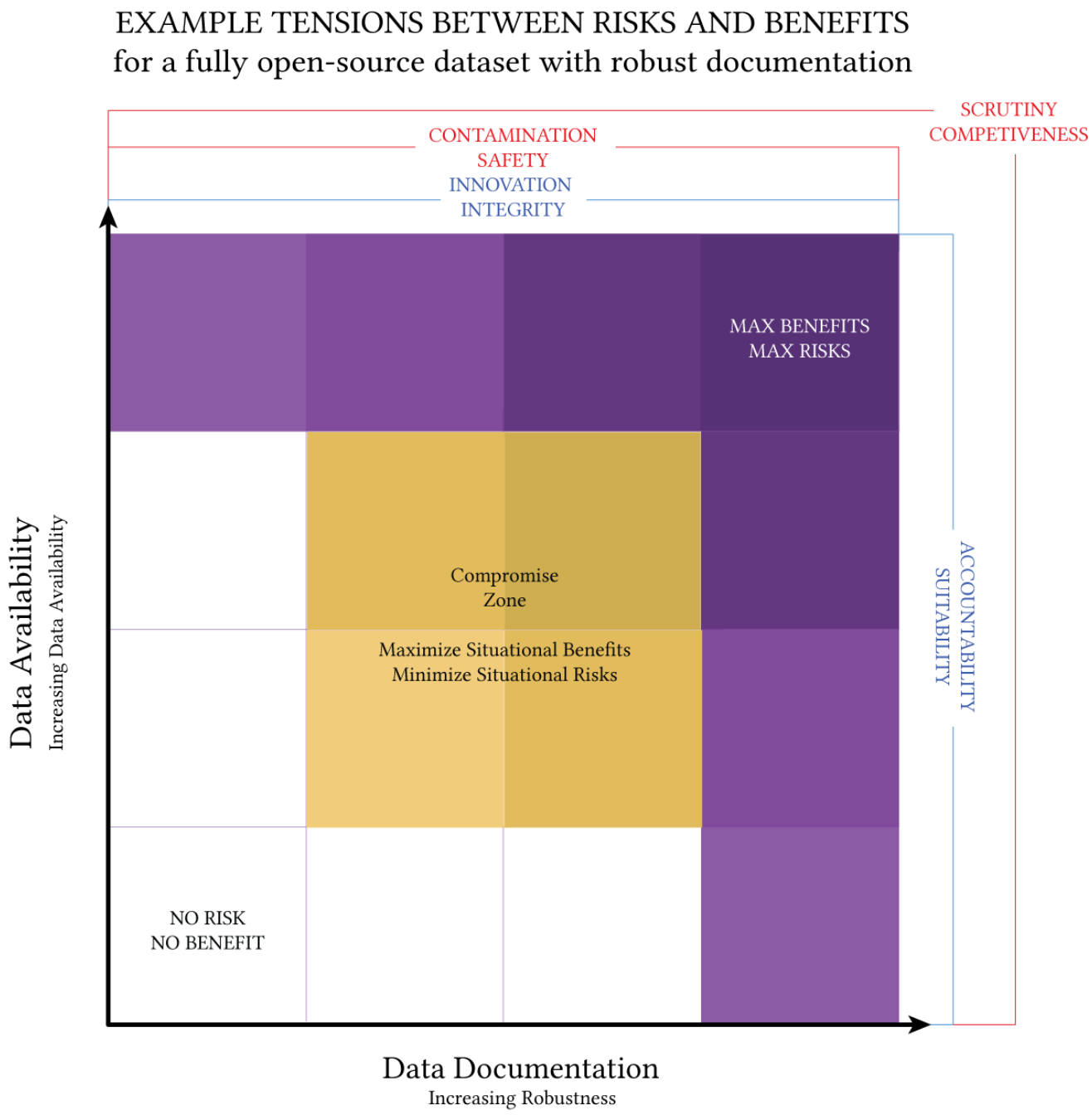



**Figure 4: A heatmap visualization that shows how both the risks and the benefits of transparency are more salient the more open availability and the more robust documentation is. Here, in our example dataset, we show how finding compromises in the middle of the heatmap may present the most maximal situational benefits and minimal situational risks.**