An Assessment of Reported Biases and Harms of Large Language Models

Heesoo Jang[1, 2] and Jaemin Cho[3]

[1] Hussman School of Journalism and Media, University of North Carolina at Chapel Hill

[2] Center for Information, Technology, and Public Life (CITAP), University of North Carolina at Chapel Hill

[3] Department of Computer Science, University of North Carolina at Chapel Hill

Note. This paper was submitted to the Human-Machine Communication Interest Group at the 74th International Communication Association (ICA) Conference, 2024. It received the Top Paper Award.

Recommended Citation. Jang, H., & Cho, J. (2024). *An assessment of reported biases and harms of large language models*. Paper presented at the 74th International Communication Association Conference, Human-Machine Communication Interest Group, Gold Coast, Australia.

Abstract

Artificial intelligence (AI) systems are perpetuating social biases, harming those who are already marginalized. In response, documenting ethical considerations of AI models has emerged as a non-algorithmic solution to assess and mitigate AI biases and harms. This study examined how biases and harms are reported and understood in the documents of so-called large language models (LLMs). We used both qualitative thematic analysis and quantitative content analysis. Based on our analysis, we discuss the implications of our findings: the need for public availability for identifying and mitigating biases, the observed consensus around understanding biases in models, bias evaluations that narrowly define bias through existing benchmarks, the need to go beyond just listing harms than discussing them, and delegation of mitigation efforts to future work and downstream applications. Our study shows that the AI industry needs more interdisciplinary collaborations with scholars who have expertise in representation, bias, prejudice, and ethics.

Keywords: AI, Algorithmic bias, Algorithmic harm, Large Language Models, NLP, Model cards, Representation harm

An Assessment of Reported Biases and Harms of Large Language Models

With artificial intelligence (AI hereafter) systems and applications widely penetrating our daily lives, there is a growing social consensus on the importance of accounting for the fairness of such systems (Mehrabi et al., 2021). Several systematic biases, including racial and gender biases, have been discovered in these applications of AI models. Commercial gender classifications APIs have error rates as high as 33% for darker-skinned females, while their performance on lighter-skinned males is near perfection (Buolamwini & Gebru, 2018). Automatic speech detection recognizes male voices better than female voices (Tatman, 2017), sentiment analysis systems rank sentences containing female noun phrases to be indicative of anger more often than those containing male noun phrases (Park et al., 2018), and image captioning models automatically predict the agent to be male if there is a computer nearby (Hendricks et al., 2018). Language models regard traditional European-American names as closer to words like joy while analyzing African-American names to be closer to words like agony (Caliskan et al., 2017).

We care about these biases and harms stemming from AI systems because of their broader impact to society. The scale and speed of the harms caused by these systems are unprecedented. The harms of these AI biases have been extensively observed and reported by several researchers. In her book *Algorithms of Oppression: How Search Engines Reinforce Racism*, Safiya Umoja Noble (2018) shows how algorithms in search engines echo the racist and sexist biases of the society, harming people of marginalized race and gender. Virginia Eubanks (2018), in her book *Automating Inequality*, investigates how algorithms used in policy reinforces class and racial biases through the data it collects from the people it is supposed to help. A more recent publication of Ruha Benjamin (2019) also reveals how algorithms contribute to White

supremacy and social inequity through how they are designed. Despite the significant amount of harm these reported biases can pose to people's lives, researchers have shown concern over the lack of information provided regarding trained models (Gebru et al., 2021; Holland et al., 2018; Mitchell et al., 2021).

In response to this societal need, documenting ethical considerations of trained machine learning (ML) models has been an emerging trend in machine learning since last year as a non-algorithmic solution to assess and mitigate AI biases and harms. Several projects have explored ways of reporting detailed performance characteristics of data and models, including the Dataset Nutrition Label project (Holland et al., 2018), Datasheets for Datasets (Gebru et al., 2021), and Model Cards (Mitchell et al., 2021). As top-tier machine learning conferences have begun to require researchers to include a discussion about potential negative societal impacts of the proposed research artifact or application (e.g., NeurIPS, 2022; ACL Rolling Review, 2022; CVPR, 2022), we expect and hope this to become a major trend in machine learning.

Documenting trained models is especially important than ever because, as Gebru and her colleagues (2021) claim, "machine learning is no longer a purely academic discipline (p. 1)." These data and trained models are applied in various domains that can have a significant impact on people's lives, including health care (e.g., Microsoft Research, 2018), employment (e.g., Strazzulla, 2022), and criminal justice (e.g., Spivack & Garvie, 2021). Both identifying and mitigating the biases and harms AI systems have become more important as the field of AI goes under a paradigm shift with the rise of pre-trained general purpose AI models, which are "trained on broad data at scale and are adaptable to a wide range of downstream tasks" (Bommasani et al., 2021, p. 1).

In this study, we focused on a specific type of AI models: “large” language models. We examined the status quo of documenting ethical considerations of these models, evaluated current practices, and provided implications and future directions for assessing and mitigating AI biases and harms. Through this project, we aimed to answer the following research questions:

RQ1. What are the kinds of information regarding biases and harms that are disclosed through documentation?

RQ2. What are the biases and harms that are generally well-documented, and what are those that are neglected or receive less attention?

RQ3. Have model cards encouraged more ethical documentation of AI biases and harms?

To do so, we first reviewed relevant literature that proposed frameworks for ethical documentation and surveys of algorithmic biases and harms. Next, we used both qualitative thematic analysis and quantitative content analysis to examine the ethical considerations reported by ten AI models, including GPT-3, Gopher, OPT, and Chinchilla, among others. Based on our analysis, we discuss the implications of our findings: the need for public availability for identifying and mitigating biases, the observed consensus around understanding biases in models, bias evaluations that narrowly define bias through existing benchmarks, the trend in listing harms than discussing them, and delegation of mitigation efforts to future work and downstream applications.

Examining documentation of the model’s ethical performances, including metrics of bias and fairness, will allow us to collect essential information about how biases and harms of large language models are assessed and mitigated. The significance of this study is twofold. First, this study provides a useful framework for assessing ethical AI reporting. Second, this study presents

a snapshot of the current state of AI ethics practices in reporting AI harm and can provide implications and directions for the future of AI ethics.

Literature Review

Large Language Models and Their Applications

Language models (LM) are computational models that approximate probability distribution over text given other text, which is often parameterized with deep neural networks. After being trained on large corpora, the language models are often adapted to diverse tasks such as speech recognition, question answering, machine translation, and information retrieval. As many researchers found that increasing the number of parameters in language models improves their performance, large language models (LLMs hereafter) have become increasingly popular.

Technically, LLMs use technologies that have been used in the AI field for decades: mainly deep neural networks and self-supervised learning (Bommasani et al., 2021). However, what makes these models significant is that they can demonstrate improved performances in a wide range of downstream tasks and even emergent capabilities through transfer learning and scale. By emergent capabilities, we refer to functions that were “neither specifically trained for nor anticipated to arise” (Bommasani et al., 2021, p.5). Most LLMs, such as BERT and GPT-3, are not directly deployed but used as intermediary assets to build domain-specific applications. For example, LLM BERT (Devlin et al., 2019) has been adapted to several domains, inspiring many variants of them: RoBERTa, XLNet, MT-DNN, SpanBERT, VisualBERT, K-BERT, HUBERT..

The range of downstream tasks to which these LLMs can be applied go beyond our imagination. A majority of the state-of-the-art language models are now adapted from only a handful of LLMs. Even more, there is a recent trend of using LLMs across a wide range of

modalities (Bommasani et al., 2021); that is, foundation models that are built for language-related tasks are also applied to images, speech, tabular data, and even protein sequences or organic molecules (examples as cited in Bommasani et al., 2021). There also have been recent advancements in making these LLMs to be multimodal themselves by training them simultaneously on both image and text datasets, thus making them no longer "language models" (e.g., CLIP, DALL-E).

Researchers and civic groups have raised concerns about the far-reaching yet unknown consequences of LLMs, which we will discuss in a latter section of this paper. As illustrated above, the significance of these models come from their capabilities in a wide range of downstream tasks, including emergent ones. As a result, it gets more complicated to grasp the societal impacts of foundation models compared to systems that have a well-specified and focused purpose (Bommasani et al., 2021). Nonetheless, the social impacts of these models are critical because the nature of these models allows quick and widespread integrations into real life.

Advocacy Towards Transparency: Public Access and Documentation

Public access. The lack of accessibility of LLMs has been criticized for hindering open science (Bommasani et al., 2021). This is a major obstacle to identifying and mitigating AI harms and biases, especially regarding emergent functionalities that are only demonstrated in models of sufficient sizes.

Before LLMs gained popularity, reproducibility and open science have been a norm within the field. Machine learning packages such as *TensorFlow* and *PyTorch* have facilitated collaborations among developers and easier access to each other's AI models. There were initiatives such as ML Reproducibility Challenge organized by Papers With Code (Papers With

Code, 2022) to foster reproducibility and major conferences adopted reproducibility checklists. Open science, such as publicly releasing both code and datasets, have been increasingly encouraged for the innovation and progress of the field (Papers with Code, 2022; Papers with Datasets, n.d.).

Nonetheless, LLMs have not followed this trend of transparency and open science. In their report, Bommasani et al. (2021) have pointed out that GPT-3 models are not released at all and even when trained models are made available (e.g., in the case of BERT), the computational cost and complex engineering requirements prevent AI researchers from having full access to these models.

Documentation. Documenting ethical considerations of trained machine learning models has been an emerging trend in the AI field since 2021 as a non-algorithmic solution to assess and mitigate AI biases and harms. Several projects have explored ways of reporting detailed performance characteristics of data and models, including the Dataset Nutrition Label project (Holland et al., 2018), Datasheets for Datasets (Gebru et al., 2021), and Model Cards (Mitchell et al., 2021). As top-tier machine learning conferences have begun to require researchers to include a discussion about potential negative societal impacts of the proposed research artifact or application (e.g., NeurIPS, 2022; ACL Rolling Review, 2022; CVPR, 2022), we expect and hope this to become a major trend in machine learning.

Model cards are currently the most widely adopted way of ethical documentation. Model cards are "short documents accompanying trained machine learning models that provide benchmarked evaluation in a variety of conditions, such as across different cultural, demographic, or phenotypic groups and intersectional groups that are relevant to the intended application domains (Mitchell et al., 2019, abstract)." Model card reporting is a framework

proposed by Mitchell et al. (2019) to encourage transparent model reporting for the following two purposes: (1) clarify the intended use cases of ML models and (2) minimize their usage in contexts for which they are not well suited. Currently, the interface of Hugging Face encourages filling out a model card.

Even when models are not public, model cards should be provided. According to Michell et al.'s (2019) proposal, model cards aim to "standardize ethical practice and reporting - allowing stakeholders to compare candidate models for deployment across not only traditional evaluation metrics but also along the axes of ethical, inclusive, and fair considerations (abstract)." Thus, even when the model itself is not publicly available, end users have the right to know the biases and harms that the model might perpetuate.

An empirical question this study will investigate is whether model cards have encouraged transparent reporting and more considerations of biases and harms of these models.

The Biases and Harms of Large Language Models

Although the term bias has been used in both the social sciences and the machine learning field, the term has been used in different contexts. In the social sciences, the term bias has been used to refer to the skews that lead to unjust discrimination based on personal traits, including but not limited to age, gender, religion, and ability status (Blodgett et al., 2020; Weidinger et al., 2019). Bias has been often used as a synonym of prejudice, which is defined as "an antipathy based on faulty and inflexible generalization" that is "directed toward a group as a whole, or toward an individual because he is a member of that group (Allport, 1954, p. 9)." On the other hand, in the machine learning field, the term bias is synonymous to 'error' and has been used to refer to "the difference between the expected (or average) prediction of our model and the correct value which we are trying to predict," also known as underfitting (Dietterich & Kong,

1995; Fortmann-Roe, 2012). It has only been recent—since AI biases have gained societal and academic attention—that the machine learning field has started using this term in the social science context, as we also observed in the results of our study.

Both terms, bias and harm, are used prevalently when discussing the societal impacts of LLMs. Despite these terms being used interchangeably, we would like to conceptually distinguish between biases and harms for the purposes of this study. A review of the literature reveals a causal relationship between bias and harm where bias is the cause and harm is the outcome of the biases (e.g., Bommasani et al., 2021; Mehrabi et al., 2019; Mitchell et al., 2019; Weidinger et al., 2019). Thus, we narrow the definition of bias to refer to "the properties of an LLM that cause harm to people and society." By suggesting the causal relationship of LLM biases and harms, we are not arguing that LLM biases are the only cause of LLM harms, but that biases precede harm.

Conceptually distinguishing LLM biases and harms is useful for the following two reasons. First, through this conceptual separation, the biases we care about becomes more clear; that is, we are able to focus on the biases that lead to harm. For example, a biased dataset may be desirable depending on the purpose of the model. A model that targets one language for future applications would want a dataset biased to that particular language. If the users served by this language only use one language, the bias would not necessarily lead to harm (e.g., developing a Korean language model to provide service in South Korea). Conversely, a monolingual model that targets a global audience could potentially lead to harm by being discriminatory to its users. Also, some biases can be observed by looking at the training dataset or the codes and checkpoints because of the intrinsic nature of biases. Yet, some biases are latent and recognized

through the harms that users experience even though they are embedded in the system at the pre-training process.

Harms can also be thought of in two ways: those that can be mitigated through resolving biases and those that should be mitigated by other sources of harm. One example of the latter is the environmental harm caused by LLMs. Moreover, harms caused by biases can be further classified as either allocative harms or representation harms (Crawford, 2017; Zhang et al., 2022a). While allocative harms refer to harms that occur when a system unfairly allocates an opportunity or resource to a specific group or people, representation harms refer to instances where a system perpetuates stereotypes and power dynamics in a way that reinforces subordination of a group.

The introduction of the concept “representation harms” to the AI ethics field changed how people think about AI biases and harms (Crawford, 2017). Previously, representation biases were discussed in terms of allocative harms. Representation biases were perceived as harmful only when they unfairly allocated opportunities and resources. However, representation harms indicate that representation biases themselves cause harm by simply existing. This argument is not new, and we will demonstrate why this is the case in the following section regarding the consequences of underrepresentation.

Hence, we can see that the terms bias and harm are conceptually different because not all biases lead to harm and not all harms stem from biases. This distinction helps to focus our effort when it comes to detecting and mitigating these biases and harms. By definition, biases are intrinsic as they are tangible byproducts of the data and the training process that the model went through. However, harms are extrinsic by nature as they occur when the model is experienced by users and not inherent in the model itself.

The Consequences of Underrepresentation

As mentioned above, representation harms are at the center of discussions regarding biases and harms of LLMs. Although representation harms have been extensively discussed in AI ethics, there have been fewer discussions on the consequences of underrepresentation and representation harms. More specifically, questions such as "what happens when people experience representation harms?" or "what are the specific consequences of harmful representations?" have been left unanswered. We believe that communication studies have a lot to offer in terms of explaining why representation harms matters and what the consequences of inaccurate representations are on people and society. Media representation studies are relevant to understanding LLM representation harms because first-hand representation harm experiences of LLMs are mediated through the media, whether through a social media platform, a chatbot service, or a website demo.

Several communication theories provide explanations on why representation harms are critical. A paradigm referred to as "media world as real world," for example, embraces the view that people process mediated experiences as real-world, first-hand experiences. Early media scholars have argued that the ways media represent both ingroup and outgroup are important because people do not experience a large proportion of the world directly, but they do so indirectly through the images in their heads that are constructed through the information from various media (Lipmann, 1922).

Similarly, the phenomenon that people do not distinguish between the mediated world and the real world has been also supported by the media equation theory (Reeves & Nass, 1996). The main argument of media equation theory is that people respond to media, communication technologies, and mediated text and images as we would to actual people and places (Reeves &

Nass, 1996). Through many experiments, Reeves & Nass (1996) and subsequent research in the HCI field have shown that people's way of feeling and making sense of the mediated world is deeply connected to those of the real world.

Evidence supports that mediated text and images provide influential exemplars in social judgment. For instance, media portrayals affect perceptions of the frequency of events such as crime and, by extension, the prevalence of crime associated with specific outgroup members (also referred to as the hostile media effect). Some studies even demonstrated that exemplars of outgroup members viewed on television had immediate intergroup attitudes even within a short amount of time after relatively little exposure (e.g., Morgan, 1982; Rossler & Brosius, 2001).

Another relevant theory is social cognitive theory (Bandura, 2002). Social cognitive theory suggests that similarity to those portrayed in media is important to learning from their behaviors. As a result, imbalanced representations can reinforce already existing inequalities that exclude and demean the value of minorities in society, implying that the dominant group are more valuable and important actors of the community (Armstrong, 2004). When the minority groups perceive that the media content do not concern or represent them, they are likely to leave the platform or service. When the medium of interest is not critical to everyday life, this may represent a business problem, where the company loses a segment of customers due to dissatisfaction. However, when the media content is relevant to information that is crucial to maintain public life or personal health, representation harms prevent minority groups from effectively participating in political communication, maintaining their civic lives, and taking care of their well-being.

Lastly, the spiral of silence theory (Noelle-Neumann, 1974) shows that people who believe that they hold a minority public opinion will remain in the background while those who

hold the majority viewpoint are more encouraged to speak up. Considering that LLMs have the capability to generate enormous amounts of text and distribute them at a fast pace, underrepresentation of minorities in these outputs have the danger to silence the voices that are already marginalized.

Method

Sample

The sample included ten language models based on the inclusion criteria that the model should have more than 20 billion parameters (See Table 1 in the Appendix).

Codes

The coding sheet consisted of three areas of analysis: ways of reporting, public availability, and biases and harms.

Ways of reporting: paper and model card. For ways of reporting, we coded whether the model had a published paper and a model card (1 = *yes*, 0 = *no*).

Public availability. Public availability included three items, which were training data, codes, and checkpoints. A model's training data is publicly available when the dataset used for training the largest model is publicly available (0 = *none*, 1 = *partially public*, and 2 = *completely public*). When it comes to codes, there are two types of codes that can be made publicly available: training codes and inference codes. Training codes refer to the codes that were used in the training process of the model, where they are used to optimize model parameters. Inference codes are the codes for performing tasks with the published model and its pre-trained parameters (also called checkpoints). Thus, codes were coded as a categorical variable, where 0 meant neither codes were available, 1 meant only training codes were available, 2 meant only inference codes were available, and 3 meant both training and inference codes were available. Lastly,

checkpoints, which refer to the set of pre-trained parameters of the model, were coded as a categorical variable: 0 = *none*, 1 = *partially public*, and 2 = *completely public*. Later, the codes item was recoded to reflect the same scale as the checkpoints items: 0 = *none*, 1 = *partially public* (meaning either the training code or the inference code was public but not both), and 2 = *completely public*. Thus, all three items included in the public availability variable eventually ranged between 0 and 3. The public availability variable was the average mean score of these three items.

Biases and harms. We first used the survey of biases in machine learning conducted by Mehrabi et al. (2019) as guidance to code and identify biases. Mehrabi's bias survey included three broad categories (data to algorithm bias, algorithm to user bias, and user to data bias), which were further divided into a total of 19 sub-categories (Table 1). While going through the documents; however, the researchers realized that most of the biases mentioned in papers and model cards all fell into the representation biases category.

Therefore, based on initial thematic coding, the researchers proposed new categories for coding biases. Focusing on representation biases, researchers coded whether the representation bias was reported at the output level (i.e., representation bias concerning the output of the model), the input level (i.e., representation bias concerning the training data of the model), and the evaluation level (i.e., representation biases concerning the data and model used for evaluating the LLM biases). Output level, input level, and evaluation level representation biases were all coded dichotomously (0 = n*ot mentioned,* 1 = *mentioned*).

Moreover, for output level representation biases, we further coded for the personal traits mentioned in regard to the bias. Our list of personal traits included race/ethnicity, sex/gender, religion, nationality, language type (including mentions of different dialects of the same

language), age, disability, sexual orientation, political ideology, socioeconomic status, and unspecified (not linking the bias to any identity and referring to it in general terms).

Harms included five categories adapted from Weidinger et al. (2019): discrimination, hate speech, and exclusion; information hazards; misinformation harms; malicious uses; and environmental harms. These categories were also coded for in cross-tabulation with the personal traits discussed in relevance to the particular bias, similarly to the coding of biases. Initially, Weidinger et al. (2019) suggested a total of six categories. However, because the sample of LLMs we analyzed were baseline language models that were not yet adopted for downstream tasks, we excluded the human-computer interaction harms category. Also, we changed the environmental and socioeconomic harms to environmental harms category to limit the scope of the category.

Lastly, bias efforts coded whether the language model went through bias evaluations and any mitigation steps. LLMs that reported both bias evaluations and any kind of mitigation step were coded as 2, only reporting bias evaluations without any mitigation steps were coded as 1, and reporting neither was coded as 0.

Procedures

The first author used thematic coding to qualitatively examine the sample and identify emerging themes. These themes functioned as the basis of searching for relevant literature. Based on the themes that emerged through the qualitative coding and the relevant literature, the researchers designed a coding sheet together, which included the variables introduced above. For training, both researchers coded Meta's BlenderBot 3 paper (Shuster, 2022), which was not included in the sample). After coding this sampler paper, the researchers went through the coding results, checked discrepancies and resolved them, and updated the coding sheet based on

discussion. Then, the researchers went on to code the documents of ten LLM papers in the sample. The researchers found no discrepancies in the coded results. As a result, this study includes both quantitative and qualitative analyses of the documentation of ten LLMs in the sample.

Result

Please refer to Table 2 in the Appendix for a summary of the results.

Ways of Reporting

90% of the sample had papers, while 60% had model cards. Model cards were either included in the paper or available separately through other platforms such as GitHub[1] or Hugging Face[2].

Public Availability

Most of the language models we surveyed had low public availability (with scores less than 1). We examined the public availability of the training data, the codes (including both the training and inference codes), and the checkpoints. We found an all-or-nothing situation for public availability; that is, half of the language models did not have any of them publicly available at all, which gave them average public availability scores of zero. Two models, BLOOM and GPT-NeoX received an average score of two, which means that these models' training data, both training and inference codes, and checkpoints were all completely public. Language model OPT was mostly publicly available, with only the checkpoints partially public[3], receiving an average score of 1.67.

Representation Bias

[1] https://github.com/
[2] https://huggingface.co/
[3] Access to the checkpoints is approved to researchers upon request for OPT-175B, the largest model (https://github.com/facebookresearch/metaseq/tree/main/projects/OPT)

We analyzed both types and sources of representation biases mentioned in the LLM documents. See table 3 in the Appendix for the summary. Almost all of the analyzed LLM documents mentioned at least one kind of representation bias in either their paper or their model cards (90%). Representation biases in race/ethnicity, sex/gender, and religion were mentioned the most (70% each). Representation biases in nationality were also mentioned in 50% of the papers. Representation biases in language types were also mentioned in 50% of the papers, referring to either their overrepresentation of the English language compared to other languages or their underrepresentation of ethnic dialects of English. Biases regarding disability, sexual orientation, and age were identities that were relatively less mentioned; only 20% of papers mentioned each of them. Political ideology and socioeconomic status were mentioned in one paper. There were also two references to representation biases without specifying any identity. None of the papers in our analysis mentioned representation biases regarding intersectional identities, although those with model cards reported that they do not explore intersectionality.

We also examined the source of the reported representation biases in the language model papers. Representation biases were most mentioned at the output level (90% of the total sample). All language model papers that mentioned representation bias talked out representation biases of the output text generated by language models. Input level representation biases were less mentioned but still appeared in 40% of the total number of papers and referred to the representation biases embedded in the text data used for training the model. Lastly, representation biases were also mentioned at the evaluation level in 40% of the total number of papers. The papers reported that their bias evaluations are subject to the representation biases embedded in the benchmark algorithms used to evaluate fairness.

Bias Efforts: Bias Evaluation and Mitigation

Quantitative bias evaluations were reported in 70% of the sample. In other words, the remaining 30% of the papers did not quantitatively examine biases in their own models. When it comes to mitigation, there was a divide on whether bias mitigations should be made at the pre-training level. 30% of the LLMs reported some kind of bias mitigation process, and these mitigations were all processed at the training data level.

Five Types of Harm

We analyzed five types of harm: 1) discrimination, hate speech, and exclusion; 2) information hazards; 3) misinformation harms; 4) malicious uses; 5) environmental harm (Please refer to Table 4 in the Appendix for a summary of reported harms). All papers included sections that discuss at least one type of harm. Harms related to discrimination, hate speech, and exclusion were mentioned in 80% of the papers, making it the most mentioned category of harms among the five types of harms we coded. Considering the papers analyzed were all LLM documents, harms related to discrimination, hate speech, and exclusion were relevant in nature as these models were susceptible to representation biases, as shown in the previous section (Result-Representation Bias). Next, misinformation harms and malicious uses were both mentioned in seven of the ten language model papers and often stated in relation to each other (e.g., the language model being used by a mal-intended user for misinformation purposes). Lastly, harms related to information hazards and environmental harms were reported in 60% of the papers. Although these two types of harm were mentioned less than other types of harm, they were still mentioned in more than half of the papers.

The Effect of Model Cards

To explore whether model cards make a difference, we compared the average mean scores of representation biases between those with model cards and those without model cards

(range: 0-3). The language models without model cards had a mean score of 1 (SD=0.82), while LLMs with model cards showed a higher mean score of 2.167 (SD=0.98). Although a sample of ten language models is too small and lacks the power to conduct an independent t-test, this result shows that language models with model cards reported representation biases from more diverse lev,els compared to those without model cards.

Discussion

More Public Availability Needed for Identifying and Mitigating biases

The public availability of codes, training data, and checkpoints is essential for identifying and examining biases and potential harms. Although making an LLM completely public comes with its own risks, public availability is a desirable value. Without publicly available models, researchers and developers—outside the organization that developed the model—lack the resources to identify and understand the biases and harms of the model at hand. Our results show that big tech corporations were unlikely to make their LLMs publicly available. These giant tech companies' monopoly over LLMs has engendered criticisms, provoking independent researchers to collaborate on building their own LLMs and making them publicly available (e.g., BLOOM by BigScience, GPT-NeoX by ElutherAI). Considering the large impact big tech companies' LLMs have on people's lives and society, these models should be more publicly available.

We also found that the decisions of whether to make the model publicly available were made by individual companies based on their own risk-benefit analyses. In other words, corporate decisions on models' public availability were not attributed to any external responsibilities or regulations and fully depended on the organization's willingness to do so. As a result, all LLMs from big tech companies except for OPT by Meta were not publicly available, and BigScience and ElutherAI were the only companies that explicitly pursued complete public

availability as a value. Recently, though, the governance frameworks suggested by Partnership on AI (2021) and NIST (Schwartz et al., 2022) have encouraged Meta to make their most recent LLM and relevant documents completely public (Zhang et al., 2022a, p. 9). In May 2022, Meta gained positive public attention by releasing their LLM OPT-175B (Zhang et al., 2022b). Meta also unprecedentedly made their logbooks (documentations of their daily training process) publicly available (Mah, 2022).

Moving Forward with Consensus Around Understanding Biases

All LLM papers that mention representation biases showed consensus on the following five claims regarding the nature of LLM biases and harms. First, these papers agreed that many LLMs inherently contain biases that harm marginalized populations by perpetuating injustice. Second, these papers commonly emphasized that downstream applications of LLMs will make visible the latent biases and harms that were not observable at the pre-training level. Third, as a result, these papers agreed that the entities that use these LLMs for downstream applications should put effort into identifying latent biases and mitigating the resulting harms. Fourth, they agreed on the limitations of the fairness benchmarks currently available. And lastly, they all underscored the importance of future research on advanced ways of identifying and mitigating these biases and harms. Overall, papers showed assent around the biases and harms of LLMs and the significance of future work.

Moreover, although not explicitly mentioned in all papers, we noticed an agreement on the importance of transparent documentation. All papers included more than a paragraph discussing ethical implications and broader impacts of LLMs, and more than half of the LLMs had publicly available model cards. The works cited in these papers demonstrate that the discussions around ethics and biases are heavily indebted to the AI ethics research community

that has examined biases in NLP (e.g., Blodgett et al., 2020; Sheng et al., 2021; Weidinger et al., 2019) and suggested new frameworks of ethical documentation (e.g., Gebru et al., 2021; Holland et al., 2020; Mitchell et al., 2019). These studies have been published in tandem with the appearance of LLMs without further delay, which led to a timely discussion of the biases and harms of LLMs.

Nonetheless, there is room for improvement. Terms such as risk, harm, bias, and toxicity were not defined when they were discussed, which repeats the observations made in previous studies (Blodgett et al., 2020; Zhang et al., 2022a). For example, biases are either undefined or loosely defined as discrimination based on demographic identity. A similar observation was also made by Sheng et al. (2021a) in their survey focusing on Natural Language Generation (NLG). Because biases are not explicitly defined, readers are left to interpret how the researchers define 'bias' through the proxy metrics they adopt to identify and measure biases.

Among the many biases that can emerge from LLMs, the sample of documents we examined mainly focused on representational biases. This tendency makes sense considering the amount of criticism LLMs have been receiving regarding the harm they do by perpetuating injustices against marginalized populations through the ways they predict language (e.g., Blodgett et al., 2020; Caliskan et al., 2017). We find a consensus that representational biases themselves are harmful in their own rights. LLM papers (and model cards) discussed representation biases regarding several types of personal traits, but none of them examined intersectionality. Since being at the intersection of several of these traits can further marginalize those who are already marginalized, we argue that intersectionality should be further examined to identify and mitigate LLM harms.

LLM documentations discussed representation biases on several levels: output level, input level, and evaluation level. By further dividing the layers of how representation biases are mentioned or discussed, this study opens the opportunity for examining how representation biases are understood and reported in LLM documents and for conducting more specific analyses of ethical documentation practices in the future.

Bias evaluations: narrowly defining bias and using existing benchmarks

While examining the way representation biases are discussed in LLM papers, we found two concerning tendencies. One tendency was to focus on the biases toward people in the text rather than biases toward people outside of the text when discussing representation bias at the output level. As mentioned in the literature review, AI biases can include those toward (a) people described in the text, (b) people who produce the text, and (c) people to whom the text is addressed (Sheng et al., 2021a). Sheng et al. (2021) have also previously reported the general tendency of AI research to narrowly define bias as bias towards people in text rather than bias towards people interacting with the model due to the relative easiness of measuring the previous definition of bias than the latter.

Another trend was less transparency on reporting about the benchmarks used to identify biases in LLMs. Researchers rarely provided the rationale for choosing one benchmark over others nor checked the benchmarks' validities. Data and information related to these benchmarks are often not publicly available, making it harder for the readers to evaluate the validity and reliability of these benchmarks. Here, we introduce two specific cases.

First was the case of using PerspectiveAPI (Perspective, n.d.) to detect toxicity. Chinchilla, PaLM, OPT, and Gopher used PerspectiveAPI either directly or indirectly via RealToxicityPrompts (which uses PerspectiveAPI) (Gehman et al., 2020). LLM papers that used

PerspectiveAPI did not explain the details of how toxicity scores should be interpreted. However, toxicity scores hold a particular meaning when LLM researchers decide to use PerpectiveAPI. A toxicity score of 0.5 does not necessarily mean a less toxic text than, for example, a score of 0.7. Rather, it means there is a discrepancy among raters on whether the text is toxic or not. The training data of PerspectiveAPI were collected in a way that the score reflects the percentage of raters that rated the text as toxic (Perspective, n.d.). If five raters found the text toxic and five others found it less toxic, this disagreement leads to a toxicity score of 0.5. Closer to the midpoint means of uncertainty; that is, the model is less confident about whether the text is toxic or not. Closer to either end of 0 and 1, the model has more confidence that the text is toxic or not. Phrases such as "greater toxicity (e.g., Rae et al., 2021, p.13)" provide the impression that the toxicity scores reflect the degree of toxicity when it actually means more confidence in toxicity. Also, the fact that PerspectiveAPI is not a publicly available model makes it hard to evaluate the validity and reliability of the benchmark. The readers have no way to resist, but accept the definitions, classifications, and ways of training that have been done on the perspective API model.

The gender-occupation bias test was another example. Among the seven LLM papers that reported efforts to identify bias in their models, five of them (Chinchilla, PaLM, Gopher, LaMDA, and GPT-3) used the gender-occupation bias test—called the Winogender test (Rudinger et al., 2018)—to measure gender bias. Winogender tests were used to see if LLMs can accurately determine the pronoun that refers to the occupation word. An ideal situation (where the LLM is unbiased) would be where the LLM correctly predicts the correct pronoun regardless of pronoun gender. Although the documentations of these LLMs cite the same research article for the method and dataset (i.e., Rudinger et al., 2018), there is a variation between the number

of occupations used by the original paper (Rudinger et al., 2018 used 60 occupations) and the number of occupations used by the LLM papers (e.g., MT-NLG used 323 occupations; Gopher and Chinchilla used 76 occupations, GPT-3 did not report). The occupation list MT-NLG used to conduct the gender-occupation bias test included pairs of gendered professions (e.g., businessman- businesswoman, nun-monk, waiter-waitress, actor-actress) and gendered terms that only include females (e.g., housewife) or males of the profession (e.g., cameraman, congressman, fisherman, handyman, councilman, policeman, patrolman, salesman, sportsman, serviceman, statesman) (See Smith et al., 2022, pp. 43-44). Because the profession is already gendered, we have less confidence that MT-NLG's gender-occupation bias test has accurately captured the gender bias of the system. We consider Gopher's occupation dataset (which Chinchilla also used) better practice because it included neutral profession terms and clearly states the sources of the terms.

In sum, using existing benchmarks left readers with a too narrow definition of bias (if any), and the validity of these benchmarks were often questionable.

Going Beyond Mentions of Harms

Moving on to harms, we found a tendency to not discuss harms in detail, merely mentioning them. One reason for merely mentioning harm seems to be the latent nature of harms of LLMs. Because harms become observable only after being experienced by people, they are reported by citing observed harms in other LLMs. As these LLMs are simultaneously applied to a range of downstream tasks, unforeseeable harms would start appearing. We propose that the organizations that developed these LLMs should constantly update their model cards to also include the harms that latently appear as the LLM gets applied downstream. This work will

require the organizations to closely follow and collaborate with the organizations that use the LLM model for downstream applications.

Delegation of Mitigation Efforts

Mitigations of LLM harms and biases seem to be delegated to future work or downstream applications along with further examinations of biases. Even though the LLM documents identify the limitations of current benchmarks, they still use them and then call for future work to examine biases that are not easily identified. Also, although these documents recognize the harms that these biases can perpetuate, they do not put in any effort to mitigate these harms. We find this trend alarming. This tendency is understandable considering that mitigation efforts work best when taken a holistic approach with robust collaborations of different communities (Stilgoe et al., 2013; Weidinger et al., 2019). Nonetheless, considering that these tech companies, which develop and publish LLMs, are the ones with the capacities and resources to develop, implement, and encourage collaborations for mitigation, these companies should take more responsibility for mitigation efforts. For example, ElutherAI –although not a tech company– explicitly encouraged researchers and developers to reach out to them if they need support with computing power to study their model (Black et al., 2022, p.11).

Conclusion

In this study, we examined how LLM documents understand and report algorithmic biases. Based on our examination of LLM documents, we argue that the AI industry needs more interdisciplinary collaborations with scholars who have expertise in representation, biases, prejudice, and ethics. Through these collaborations, we expect clearer conceptual and operational definitions of biases and harms that could solve the construct validity problems that the current documents seem to be facing.

This study is not without limitations. For this study, we only investigated the documentation of ten LLMs. Thus, the findings of this study cannot be generalized beyond the sample. Nonetheless, because the sample included all LLMs in the inclusion criteria, the results still capture the general picture of biases and harms discussed in the LLM context. To further assess the current state of ethical documentation as a larger trend in ML and AI fields as a whole, more AI models should be examined on how they understand and report biases and harms.

References

Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., ... & Zheng, X. (2016). TensorFlow: a system for Large-Scale machine learning. In *12th USENIX symposium on operating systems design and implementation (OSDI 16)* (pp. 265-283).

ACL Rolling Review. (2022). Call for Papers. Retrieved from https://aclrollingreview.org/cfp

Allport, G. W. (1954). The Nature of Prejudice. Cambridge, MA: Addison-Wesley.

Armstrong, C. L. (2004). The influence of reporter gender on source selection in newspaper stories. *Journalism & Mass Communication Quarterly*, *81*(1), 139-154.

Benjamin, R. (2019). *Race after technology: Abolitionist tools for the new jim code.* Social forces.

BigScience. (2022, Jul 6). BLOOM. *Hugging Face*. Retrieved from https://huggingface.co/bigscience/bloom

Black, S., Biderman, S., Hallahan, E., Anthony, Q., Gao, L., Golding, L., ... & Weinbach, S. (2022). Gpt-neox-20b: An open-source autoregressive language model. In *Proceedings of the Workshop on Challenges & Perspectives in Creating Large Language Models*, pages 95–136, virtual+Dublin. Association for Computational Linguistics.

Blodgett, S. L., Barocas, S., Daumé III, H., & Wallach, H. (2020). Language (technology) is power: A critical survey of" bias" in nlp. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5454–5476, Online. Association for Computational Linguistics.

Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., ... & Liang, P. (2021). On the opportunities and risks of foundation models. arXiv preprint arXiv:2108.07258.

Caliskan, A., Bryson, J. J., & Narayanan, A. (2017). Semantics derived automatically from language corpora contain human-like biases. *Science*, *356*(6334), 183-186.

Crawford, K. (2017, Dec 10). The Trouble with Bias. Presented as Keynote at the *Thirty-first Annual Conference on Neural Information Processing Systems (NIPS)*. Retrieved from https://www.youtube.com/watch?v=fMym_BKWQzk

CVPR. (2022). Ethics Guidelines. Retrieved from https://cvpr2022.thecvf.com/ethics-guidelines

Dietterich, T. G., & Kong, E. B. (1995). *Machine learning bias, statistical bias, and statistical variance of decision tree algorithms.* Technical report, Department of Computer Science, Oregon State University.

Eubanks, V. (2018). *Automating inequality: How high-tech tools profile, police, and punish the poor*. St. Martin's Press.

Fortmann-Roe, S. (2012, June.). Understanding the Bias-Variance Tradeoff. Retrieved from http://scott.fortmann-roe.com/docs/BiasVariance.html

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

Gehman, S., Gururangan, S., Sap, M., Choi, Y., & Smith, N. A. (2020). Realtoxicityprompts: Evaluating neural toxic degeneration in language models. Presented at *the Association for Computational Linguistics: EMNLP 2020,* pp. 3356–3369.

Guest, G., MacQueen, K. M., & Namey, E. E. (2011). Applied thematic analysis. Sage publications.

Hendricks, L. A., Burns, K., Saenko, K., Darrell, T., & Rohrbach, A. (2018). Women also snowboard: Overcoming bias in captioning models. In *Proceedings of the European Conference on Computer Vision (ECCV),* pp. 771-787.

Holland, S., Hosny, A., Newman, S., Joseph, J., & Chmielinski, K. (2018) The dataset nutrition label: A framework to drive higher data quality standards. *CoRR*. Abs/1805.03677

Lippmann, W. (1922). *Public Opinion*. New York, Macmillan.

Mah, P. (May 11, 2022). Meta AI Giving Away Its New Large Language Model. *CDO Trends*. Retrieved from https://www.cdotrends.com/story/16435/meta-ai-giving-away-its-new-large-language-model

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A survey on bias and fairness in machine learning. *ACM Computing Surveys (CSUR)*, *54*(6), 1-35.

Microsoft Research. (n.d.). Project InnerEye – Medical Imaging AI to empower Clinicians. Accessed February 28, 2022 at https://www.microsoft.com/en-us/research/project/medical-image-analysis/

Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., … & Gebru, T. (2019, January). Model cards for model reporting. In *Proceedings of the conference on fairness, accountability, and transparency* (pp. 220-229).

Morgan, M. (1982). Television and adolescents" sex role stereotypes: A longitudinal study. Journal of Personality and Social Psychology, 43, 947–955.

NeurIPS. (2022). Ethics Guidelines. Retrieved from https://neurips.cc/public/EthicsGuidelines

Noble, S. U. (2018). *Algorithms of oppression*. New York University Press.

Noelle Neumann, E. (1974). The spiral of silence a theory of public opinion. *Journal of communication*, *24*(2), 43-51.

Papers with code. (2022). ML Reproducibility Challenge 2022. Retrieved from https://paperswithcode.com/rc2022

Papers with datasets. (n.d.). Retrieved from https://paperswithcode.com/datasets

Park, J. H., Shin, J., & Fung, P. (2018). Reducing gender bias in abusive language detection. arXiv preprint arXiv:1808.07231.

Partnership on AI. (2021, May 6). Six Recommendations for Responsible Publication. Retrieved from https://partnershiponai.org/paper/responsible-publication-recommendations/

Perspective. (n.d.) Perspective API. Retrieved from https://perspectiveapi.com/

Rae, J. W., Borgeaud, S., Cai, T., Millican, K., Hoffmann, J., Song, F., ... & Irving, G. (2021). Scaling language models: Methods, analysis & insights from training gopher. *arXiv preprint arXiv:2112.11446*.

Reeves, B., & Nass, C. (1996). The media equation: How people treat computers, television, and new media like real people. *Cambridge, UK*, *10*, 236605.

Rössler, P., & Brosius, H. B. (2001). Do talk shows cultivate adolescents' views of the world? A prolonged-exposure experiment. *Journal of communication*, *51*(1), 143-163.

Rudinger, R., Naradowsky, J., Leonard, B., & Van Durme, B. (2018). Gender bias in coreference resolution. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 8–14, New Orleans, Louisiana. Association for Computational Linguistics.

Schwartz, R., Vassilev, Ap., Greene, K., Perine, L., Burt, A., & Hall, P. (2022, March). *Towards a standard for identifying and managing bias in artificial intelligence*. National Institute of Standards and Technology (NIST), U.S. Department of Commerce. Retreived from https://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.1270.pdf

Sheng, E., Chang, K. W., Natarajan, P., & Peng, N. (2021). Societal biases in language generation: Progress and challenges. In *Proceedings of the 59th Annual Meeting of the*

Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pages 4275–4293, Online. Association for Computational Linguistics.

Shuster, K., Xu, J., Komeili, M., Ju, D., Smith, E. M., Roller, S., ... & Weston, J. (2022). BlenderBot 3: a deployed conversational agent that continually learns to responsibly engage. *arXiv preprint arXiv:2208.03188*.

Smith, S., Patwary, M., Norick, B., LeGresley, P., Rajbhandari, S., Casper, J., ... & Catanzaro, B. (2022). Using deepspeed and megatron to train megatron-turing nlg 530b, a large-scale generative language model. *arXiv preprint arXiv:2201.11990*.

Spivack, J., & Garvie, C. (2020). A taxonomy of legislative approaches to face recognition in the United States. Regulating Biometrics: Global Approaches and Urgent Questions, 86-95.

Strazzulla, P. (2022, Jan 18). The Top 12 Best AI Recruiting Tools – 2022. SelectSoftware Reviews. Retrieved from https://www.selectsoftwarereviews.com/buyer-guide/ai-recruiting

Weidinger, L., Mellor, J., Rauh, M., Griffin, C., Uesato, J., Huang, P. S., ... & Gabriel, I. (2021). Ethical and social risks of harm from language models. *arXiv preprint arXiv:2112.04359*.

Zhang, S., Roller, S., Goyal, N., Artetxe, M., Chen, M., Chen, S., ... & Zettlemoyer, L. (2022a). Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*.

Zhang, S., Diab, M., & Zettlemoyer, L. (2022b, May 3). Democratizing access to large-scale language models with OPT-175B. *Meta AI*. Retrieved from https://ai.facebook.com/blog/democratizing-access-to-large-scale-language-models-with-opt-175b/

Appendix

Table 1

A Summary of the Large Language Models Included in this Study

Model name	Year released	Company	Parameters (#)
Chinchilla	2022	Deepmind	70B
BLOOM	2022	BigScience	176B
GPT-NeoX	2022	ElutherAI	20B
PaLM	2022	Google	540B
OPT	2022	Meta	175B
Gopher	2021	Deepmind	280B
HyperCLOVA	2021	Naver	82B
MT-NLG	2021	Nvidia, Microsoft	530B
LaMDA	2021	Google	137B
GPT-3	2020	OpenAI	175B

Table 2

The Public Availability of LLMs in This Study

Model name	Year released	Company	Number of Parameters	Paper (1=Yes, 0=No)	Model Card (1=Yes, 0=No)	Public Availability (0 – 3) [a)]

Chinchilla	2022	Deepmind	70B	1	1	0
BLOOM	2022	BigScience [b]	176B	0	1	2
GPT-NeoX	2022	ElutherAI [b]	20B	1	0	2
PaLM	2022	Google	540B	1	1	0
OPT	2022	Meta	175B	1	1	1.67
Gopher	2021	Deepmind	280B	1	1	0
HyperCLOVA	2021	Naver	82B	1	0	0
MT-NLG	2021	Nvidia, Microsoft	530B	1	0	0.67
LaMDA	2021	Google	137B	1	0	0
GPT-3	2020	OpenAI	175B	1	1	0.33

a) See Codes - Public availability for scoring details

b) BigSceince (https://bigscience.huggingface.co/) and ElutherAI (https://www.eleuther.ai/about/) are open collaborations of independent researchers

Table 3

A Summary of Reported Biases and Bias Efforts

Model name	Representation Bias	Bias Effort

	Input level	Output level	Evaluation level	Total	Evaluation	Mitigation	Total
Chinchilla	1	1	0	2	1	0	1
BLOOM	0	1	0	1	0	0	0
GPT-NeoX	0	0	0	0	0	0	0
PaLM	1	1	1	3	1	1	2
OPT	0	1	0	1	1	0	1
Gopher	1	1	1	3	1	1	2
HyperCLOVA	0	1	0	1	0	0	0
MT-NLG	0	1	0	1	1	0	1
LaMDA	0	1	1	2	1	1	2
GPT-3	1	1	1	3	1	0	1
Total	4	9	4	9	7	3	10

Table 4

Different Types of Harm Reported in LLM Documentations

Model Name	Discrimination, Hate speech, and Exclusion	Information Hazards	Misinformation Harms	Malicious Uses	Environmental Harm
Chinchilla	1	1	1	1	0
BLOOM	1	1	1	1	0

GPT-NeoX	1	0	1	0	1
PaLM	1	1	0	1	1
OPT	1	0	1	0	1
Gopher	1	1	1	1	0
HyperCLOVA	1	1	0	1	1
MT-NLG	0	0	0	0	0
LaMDA	1	1	1	1	1
GPT-3	0	0	1	1	1
Total	8	6	7	7	6