SoUnD Framework: Analyzing (So)cial Representation in (Un)structured (D)ata

Mark Díaz, Sunipa Dev, Emily Reif, Remi Denton, Vinodkumar Prabhakaran

Google Research
markdiaz@google.com, sunipadev@google.com, ereif@google.com, dentone@google.com, vinodkpg@google.com

Abstract

Decisions about how to responsibly collect, use and document data often rely upon understanding how people are represented in data. Yet, the unlabeled nature and scale of data used in foundation model development poses a direct challenge to systematic analyses of downstream risks, such as representational harms. We provide a framework designed to help RAI practitioners more easily plan and structure analyses of how people are represented in unstructured data and identify downstream risks. The framework is organized into groups of analyses that map to 3 basic questions: 1) Who is represented in the data, 2) What content is in the data, and 3) How are the two associated. We use the framework to analyze human representation in two commonly used datasets: the Common Crawl web corpus (C4) of 356 billion tokens, and the LAION-400M dataset of 400 million text-image pairs, both developed in the English language. We illustrate how the framework informs action steps for hypothetical teams faced with data use, development, and documentation decisions. Ultimately, the framework structures human representation analyses and maps out analysis planning considerations, goals, and risk mitigation actions at different stages of dataset and model development.

Introduction

Data is widely recognized as a core underlying factor contributing to machine learning model behaviours that are potentially unfair or harmful to humans (Paullada et al. 2021). For researchers in Responsible AI (RAI), understanding the representation of social groups in data is an integral step toward identifying and mitigating social risks of AI. Indeed, Annex XI of the EU AI Act explicitly details technical documentation that providers of General-Purpose AI models must make available, including the provenance and main characteristics of data points and "measures to detect the unsuitability of data sources and methods to detect identifiable biases (Kop 2021)." Principled analysis of data underpinning pre-trained foundation models is particularly salient given their increasing reach and their use by researchers and developers who lack the resources to develop computationally-intensive models (Bommasani et al. 2021; Han et al. 2021).

At the same time, the large, unstructured nature of datasets that underpin foundation models poses significant challenges for conducting analyses required to make development, documentation, and use decisions. The open-ended potential for downstream use, means that risks are wide-ranging and sometimes lack clear methods of evaluation (Weidinger et al. 2021), complicating systematic study. Prior systematic fairness audits and auditing tools have often focused on labeled datasets and utilized aggregated and disaggregated analyses to identify class imbalances (e.g., (Saleiro et al. 2018; Kearns et al. 2018; Kleinberg, Mullainathan, and Raghavan 2016; Friedler et al. 2019)). Despite increased scrutiny of large unstructured datasets (Birhane, Prabhu, and Kahembwe 2021; Dodge et al. 2021), methods of analysis remain less robust and less systematic relative to labeled datasets, in part because labels provide a crucial pointer to dataset features to evaluate for fairness and bias concerns. Indeed, RAI practitioners have noted a dearth of standardized approaches to conducting 'fairness-aware' data collection and evaluation (Holstein et al. 2019). As a consequence, it is more difficult to make risk mitigation decisions and properly develop transparency artifacts.

Our work closes this gap by contributing a conceptual framework to facilitate standardized analysis of unstructured data in service of responsible data use and dataset development, use, and documentation. The framework pulls from published research in ML fairness and auditing to focus on how people are represented in data, including the range of data features that indicate social identity and influence the representation of different social and cultural groups– from identity terms that explicitly name social groups to recurring image objects implicitly associated with those groups. Analyses are grouped according to *who* is in the data (e.g., social identity terms, dialects, skin tones, voice pitches, etc), *what* is in the data (e.g., topics, sexual or violent content, source geography, etc), as well as associations between them. In this paper we focus on the framework's conceptual structure and point readers to the more detailed list of operationalized analyses[1]. The framework is designed to integrate into RAI workflows to support decision making for mitigation steps (e.g., data filtering, rebalancing) for a range of goals, such as new dataset development, existing dataset adaptation, and benchmark development. As a result, the framework can support a range of users and goals. Three primary use cases

[1] https://arxiv.org/abs/2311.17259

we foresee include development teams analyzing training data to assess the fairness of candidate datasets or mixtures, researchers structuring studies on the downstream impacts of data mitigations, and product teams planning compliance audits. The framework's focus on whether and how people are depicted means that the core analytical questions are not modality-specific and are extensible to new modalities and combinations as they emerge.

Background

Dataset Transparency and Documentation

Transparency is a core focus in RAI, with a growing body of scholarship aimed at increasing transparency of AI systems and datasets for a variety of stakeholders. These range from developers who may build on pre-trained models to individuals who may be subject to algorithmic decision making (Lima et al. 2022; Wagner et al. 2020). Calls for transparency have led to large-scale auditing efforts such as (Longpre et al. 2023a)'s audit of the licensing and use of over 1800 datasets and (Bommasani et al. 2023)'s index for measuring foundation model transparency. At the dataset level, transparency serves to highlight critical information both about the contents of a dataset as well as the processes that underpin how a dataset was created. To this end, a range of work brings structured approaches to documenting both dataset content and development processes (Bender and Friedman 2018; Gebru et al. 2021; Dodge et al. 2021; Díaz et al. 2022; Rostamzadeh et al. 2022; Srinivasan et al. 2021; Pushkarna, Zaldivar, and Kjartansson 2022).

However as massive, unstructured datasets increasingly become the norm in ML development, structured frameworks are needed to help summarize key characteristics of the data they capture. Dodge et al. (2021) offer an expansive audit of C4 (Raffel et al. 2020), investigating its metadata, its contents, as well as filtered data it excludes based on blocklists. Their work inspires a more structured approach to dissecting the contents of web-crawled data that feature heavily in ML datasets. Indeed, similar audits are starting to be used in foundation model development. For example, (Chowdhery et al. 2022) conducted analyses of gender, race, and religious representation in underlying training data, and Elazar et al. (2023) developed an open source platform to run standardized analyses aimed at comparing large text datasets. Such tools are invaluable and there is an opportunity to further standardize approaches, particularly for providing insights into sociotechnical risks and harms. Our work aims to standardize approaches to support existing transparency and documentation efforts by enabling the identification and communication of potential social risks associated with data.

Dataset Audits

Datasets underpinning ML training and testing have been at the center of a range of ethical controversies relating to privacy, consent, unfair system performance, representational harms, and harmful applications (Paullada et al. 2021). Against this backdrop, prominent ML datasets have been subject to close scrutiny, with empirical examinations uncovering a range of problematic content that itself is a cause of harm (e.g., copyright violations; representational harms such as misgendering) or that can lead to downstream harms, such as personally identifiable information (PII) vulnerable to model attack methods (Carlini et al. 2021). In the ML fairnesss community, many scholars have conducted dataset and model audits for social biases as well as developing new techniques to detect them. To this end, audits have focused both on representational harms (Barocas et al. 2017; Blodgett et al. 2020), as well as investigations of sensitive content, such as PII, which may be universally sensitive, but which have differential impacts for stigmatized or over-surveilled groups (Hutchinson et al. 2020).

Both image and text datasets have been shown to contain co-occurrence statistics that mirror harmful social biases and stereotypes (Garg et al. 2018; Hendricks et al. 2018); image datasets have been found to include problematic sexual imagery, including depictions of sexual violence and non-consensual sexual content, and racial and ethnic slurs within image labels and captions (Birhane and Prabhu 2021; Birhane, Prabhu, and Kahembwe 2021); text datasets have been found to contain biased and harmful sentiments towards marginalized identity groups (Hutchinson et al. 2020; Dodge et al. 2021) and to exclude perspectives from marginalized groups (Dodge et al. 2021). Dataset audits can close documentation gaps (Dodge et al. 2021) or be used to make data filtering or re-balancing decisions (Russakovsky et al. 2014). In some extreme cases, audits have led to the deprecation of datasets, such as MegaFace (Kemelmacher-Shlizerman et al. 2016), Tiny Images (Torralba, Fergus, and Freeman 2008), and, most recently, LAION-5B which was found to contain child sexual abuse material (Thiel 2023). Organizing prior individual audits, we present a principled framework that supports dataset auditing to shape dataset and model development decisions.

Standardizing Responsible AI Workflows

Evaluating datasets in a structured way and effectively communicating results to various stakeholders remains an important challenge for RAI. As Sambasivan et al. (2021b) show, data work often takes a backseat to work focused on developing state of the art models and algorithms. In addition to lower incentive to engage in data work as a component of robust evaluation, current approaches to data documentation are "largely ad hoc and myopic in nature" (Heger et al. 2022) and practitioners face difficulty in understanding why documentation is needed, how best to document, and, ultimately, what to document (Chang and Custis 2022). A range of development toolkits and checklists have been proposed to address these challenges, including documentation frameworks such as Data and Model Cards (Gebru et al. 2021; Mitchell et al. 2019; Pushkarna, Zaldivar, and Kjartansson 2022), internal auditing frameworks (Raji et al. 2020), analysis tooling such as Know Your Data[2] and WIMBD (Elazar et al. 2023), and impact assessment frameworks (Schiff et al. 2020).

For RAI audits in particular, there is a need to structure evaluations to best support decision making toward mitigat-

[2]https://knowyourdata.withgoogle.com/

ing downstream social risks. Part of the challenge of running dataset audits is determining what to measure and how to measure it. Sambasivan et al. (2021b) make this point very explicitly in connection to risk that can compound as a result of poorly evaluated data. In addition to dataset research that explicitly contributes to scientific understandings of the nature of bias, RAI is faced with a pragmatic need to support practitioners who require support in determining how generally known biases may emerge in their own data. Mitchell et al. (2022) give a high-level framework for measuring large, unstructured datasets and we extend this by configuring our framework around social representation and demonstrate how the results of an audit might be used to distill dataset decisions.

Framework

In this section, we introduce the conceptual framework for systematic evaluation, anchoring on risk and harm associated with representations of humans in data. We limit our discussion to conceptual aspects of the framework, which is further operationalized and detailed on ArXiv[3]. The framework supports analyses for a variety of goals ranging from dataset development to third-party audits. In practice, we envision three general use cases for the framework: 1) development teams evaluating pre-training data sources or mixtures for fairness concerns in dataset development and use, 2) development teams planning and carrying out audits for compliance and documentation, and 3) researchers structuring studies to understand the downstream impacts of data and model mitigations. Our framework also identifies a set of components that guide the operationalization of each analysis and the interpretation of their results.

Framework Analyses

Figure 1 demonstrates the framework's conceptual structure. We organize the framework around high-level questions about human-centered considerations in data: namely, *Who* is in the data, *What* is in the data, and *How* are the two associated? This structure also allows the analyses to focus on data questions at different levels of complexity with respect to corresponding downstream harms, as well as to prevent an over-focus on optimizing isolated analyses or metrics. Each analysis listed in the framework is accompanied by descriptions of relevant research in ML fairness

Rather than propose better optimized or more comprehensive individual analyses, we provide a structure to organize evaluations of social risk in data. Thus, analysis groupings must be operationalized according to the data modality to which the framework is applied. While selecting and applying evaluations or audit approaches from published literature to analyze and document data may seem a straightforward task, this data work faces underinvestment and limited structure (Sambasivan et al. 2021b; Heger et al. 2022). We draw from published work on biases in text, image, and image-text data as references for developing and describing the framework, but the framework is adaptable to other modalities with appropriate changes. For instance, social characteristics that appear in images may not be present at all in text data, such as with skin tone, or may be measured through completely different means, such as measuring gender representation using image classifiers rather than identity term lists for text data. Analyses can also be modified, added, or removed as the field's collective sociotechnical understanding about relevant social biases evolve over time, while preserving the overall framework structure. For example, salient social identity term lists may iteratively change as best practices respond to social shifts, or as global sociocultural contexts are increasingly integrated into RAI considerations. Analyses can also be updated alongside our understanding of salient social risks, the human social characteristics they are connected to, and our technical means of analyzing them. However, the motivating questions and risk mitigation goals remain stable.

Who is in the data? In asking who is in the data, we consider several human factors of data that include identifying the presence of people and social characteristics.

Presence of People: The analyses of people's presence specifically tally whether individuals or identifying information appear in data. This includes calculations of personally-identifiable information and face or person detection. Extending to new modalities, the analyses implicitly ask which data characteristics indicate the presence of people or might be recognized as a person, such as faces or bodies in visual data or voice in audio data. The results of these analyses are intended to guide more focused, follow up analyses that assess depictions of social groups.

Social Characteristics: The analyses of social characteristics in data center on data features that are often associated with or used as proxies for social identity. Some proxies appear directly in data, such as pronouns, while others, such as perceived age or gender expression in images, must be inferred, frequently using predictive methods (e.g., Lanitis, Draganova, and Christodoulou (2004)). Other social characteristics in data include dialect, linguistic style, skin tone, and voice pitch. These analyses provide insight into over- and under-representation of specific social groups, which has been associated with disparities in performance (Wilson, Hoffman, and Morgenstern 2019; Buolamwini and Gebru 2018) and general problems for class prediction (Johnson and Khoshgoftaar 2019). Because these characteristics are social in nature, their measurement must be localized. For example, social identity terms and gender presentation vary across social and cultural contexts and, thus, rely on curated term lists or localized references to measure.

What is in the data? The second high-level grouping of analyses focuses on content that may heighten the sensitivity of human representation as a whole.

Content: This group of analyses is focused on content characteristics that relate to harmful or undesirable outcomes that are independent of specific people or social groups. This includes sexual content, violent content, offensive content, and any subject matter that may be legally or socially sensitive. In text, for example, subject matter can be measured through topic distribution, which provides a bird's-eye view of the composition of the data including sex-

[3]https://arxiv.org/abs/2311.17259

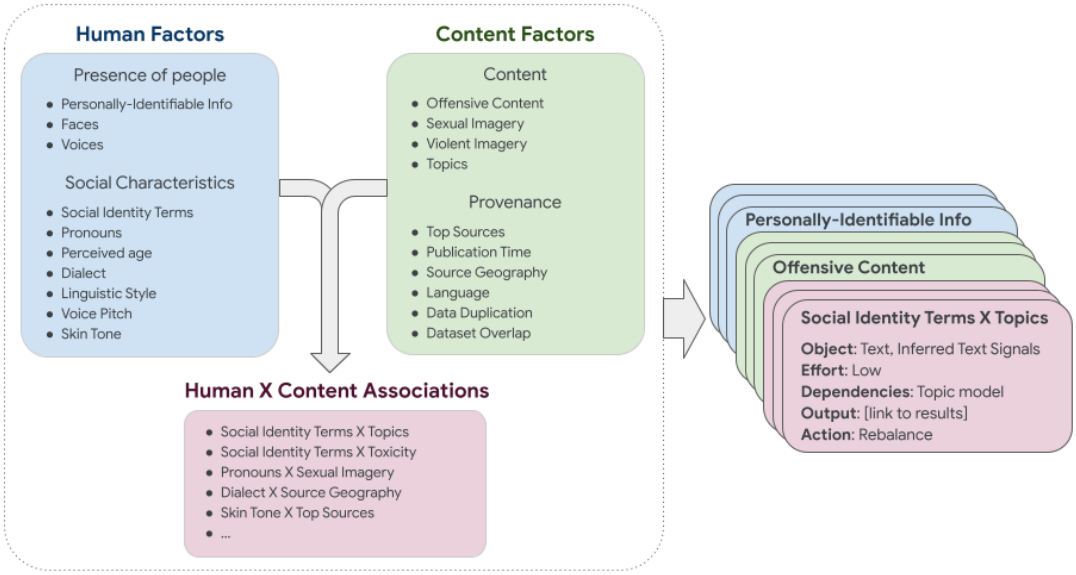


Figure 1: Framework conceptual structure. The combination of Human Factor and Content Factor analyses provided disaggregated associations. Analyses documentation and results are on the right.

ually explicit or sensitive topics. Different kinds of sensitive imagery can be assessed in visual data using specialized classifiers. Aside from obviously sensitive subject matter, content distributions can give clues to subtle downstream biases. For example, models trained primarily on news data have been shown to exhibit biases against particular country names and professions (Huang et al. 2019).

Provenance: Data provenance indicates values, norms, and perspectives in data ascertained through metadata, such as the geographic distribution of sources or publication dates. For example, source URL domains point to the range of content represented in web-scraped data, which offers insight into document content, such as linguistic and cultural content, as well as the prevalence of machine-generated text (Dodge et al. 2021). The geographic, cultural, and social representation in data can have implications for downstream models. For example, image classifiers trained on datasets sourced predominantly from western countries have lower rates of accuracy when applied to images from non-western countries (Shankar et al. 2017). Data recency also influences model performance, such as with low-resource language support, which can rely on religious or historical texts due to data scarcity (e.g., Ahmadi and Masoud (2020)).

Human $\times$ Content Associations Each standalone section of the framework can be used to provide compositional snapshots of a dataset, however, associations reveal how subgroups of people are differentially represented in data. Associations disaggregate analyses within and across modalities, such as social identity terms and topics in text or occurrences between objects detected in images and identity terms in associated text (for multimodal datasets). Associations can reveal stereotype-aligned correlations, which can propagate exclusionary norms (Dev et al. 2021a; Weidinger et al. 2021; Zhao et al. 2018; Hendricks et al. 2018).

Framework Components

Next, we outline additional framework components that guide analysis results reporting and general analysis planning. The **Output** and **Action** fields are provided to capture the results of a given analysis and any mitigation actions decided in response. Section discusses in more depth the process of making mitigation decisions. In addition to a research-backed motivation related to downstream risks, each analysis includes additional fields to support planning:

Analysis Object indicates whether an analysis is calculated on data directly (i.e., tokens in text data) or if an analysis applies to an inference produced by an intermediate classifier (e.g., inferred document topic; predicted age of person in an image). This distinction highlights which analyses are dependent on predictive models and therefore susceptible to biases that those predictive models may themselves exhibit. The distinction between "Image" and "Inferred image signals" is particularly important since few of the analyses in the framework are applied to image data directly.

Effort indicates estimated time and cost of an analysis based on current techniques and available tooling, which, in turn, reflect bias toward use for English language and Western data. In non-Western contexts, effort is likely higher for technical implementation as well as making determinations about which identity groups should be prioritized.

Dependencies indicates intermediate resources needed to conduct an analysis, such as identity term lists or classifiers which produce inferred signals. While the framework does not dictate a required implementation for any analysis, we

Analysis Goals	
Dataset Development	Developing a dataset for training or evaluation through new data collection and/or adaptation of existing datasets
Use Decisions	Making decisions regarding appropriate use of a dataset, whether for training or evaluative purposes
Model Understanding	Investigating potential roots of or explanations for model behavior
Auditing	Auditing a dataset, whether third-party or internal to a development project, for example to fill documentation gaps, ensure legal or institutional compliance, or to foster greater public awareness

Table 1: A non-exhaustive list of data analysis goals.

point to example classifiers and term lists that may be used.

Taking Action

Scoping Analyses

A first task in using the framework lies in scoping the analyses to be run. The framework is not meant to be exhaustively implemented, since the cross product of all possible Human and Content analyses would produce an intractably large number of results– all of which may not be relevant or equally impactful for the task, data, or social group in question. Similarly, while highly specific combinations of analyses can be run (e.g., an evaluation of queer depictions in Spanish-language medical literature from a specific time span), users of the framework are intended to begin with the most general question (i.e., Are people depicted in this data?; What proportion of data features sensitive content?) followed by more specific inquiries (e.g., At what rate are different groups depicted alongside sensitive content?) to provide a tractable entry point for RAI reporting and mitigation. To aid in identifying a reasonable starting point, we demarcate priority analyses that focus on analyzing age, gender, and racial/ethnic identities and their associations with sensitive content. These analyses provide a robust starting point for understanding representational harms in data.

Although the framework can be used to uncover previously unanticipated biases, in practice, users will likely already have aims informed by institutional policies regarding data use, familiarity with established benchmark evaluations, or research goals centered on particular risks. Beyond calculations of representational distributions, the framework scaffolds follow up analyses to better understand disparities and to investigate whether data distributions reflect problematic outcomes in model evaluations.

Selecting Mitigations

The most appropriate mitigation actions depend highly on context: the downstream effects of dataset risk mitigations are, in many cases, still an open question, and there are a range of other actions that may be applicable. Critically, in the development of foundation models, in particular, the cost and resources required for development pose particular challenges for studying the effects of mitigations. As a result, the effects of data mitigations remain an area of needed research. An exhaustive review of current risk mitigation approaches is beyond the scope of this work, however we describe key considerations that inform the actions a user should take. We also include a selection of guiding questions and considerations at the start of the full framework.

Important considerations are 1) dataset purpose (e.g., training or evaluation data) 2) analysis goals (e.g., auditing a third-party dataset, developing a training dataset, or developing a new evaluation benchmark), and 3) development steps at which interventions can be applied (e.g., data collection, model evaluation, documentation).

Dataset Purpose: The framework can be applied to a range of dataset types including pre-training or fine-tuning datasets. It can also be used for understanding and evaluating model-generated data. Suitable framework actions depend on the purpose of the dataset. For example, when analyzing pre-training data, it may be unclear how changes to data distributions will impact model performance, potentially making other mitigations more desirable. In contrast, data used for benchmark development stands to be used as a repeated measure of model robustness and performance. Thus, actions that might require additional costs or resources may be more easily justified to meet evaluation goals.

Analysis Goals: A range of goals can motivate framework use– each of which brings attention to different actions. Table 1 lists common goals of dataset analyses. For example, developing a new dataset from web-scraped data raises potential decisions to collect additional data or adjust filtering criteria in the data collection process. In contrast, conducting an audit of a third-party dataset for compliance purposes brings focus to documentation and use decisions.

Development Phase: Each development phase affords different actions to address data concerns. Table 2 shows common actions by development phase. For example, during data collection, toxic content biased across social identity groups might be addressed by modifying the dataset or by adjusting model evaluation planning. During the documentation phase, results of dataset analyses inform guidance on usage and flag concerns for public consumption. Data concerns may be addressed through data release decision such through licensing or access policies.

The decision to pursue an action such as the ones listed above will depend on analysis goals, available resource play a key role in determining the mitigation actions that are available. Direct action on a dataset may not be possible, for example if there are cost constraints or if data sources and filtering techniques used to develop a third-party dataset are not clearly defined or known.

Applying the Framework

In order to meet practitioner challenges and showcase how the framework guides evaluations, we provide two toy demonstrations using C4 (Raffel et al. 2020) and LAION-400M (Schuhmann et al. 2021)– two large, unstructured

Development Phase	Actions	Description
Data Collection/ Processing	Addition	Rebalancing distributions across an entire dataset or within specified categories with additional (potentially synthetic) data
	Removal	Filtering data to remove unwanted content
	Augmentation	Augmenting data, such as through data tagging (Anil et al. 2023) to allow a model to learn undesirable content while controlling its production downstream
	Flagging	Flagging analysis results for further downstream evaluation or documentation
	Non-Use	Not using the dataset, for example if applying analyses to different candidate datasets to decide which to use
Model Evaluation	Addt'l Benchmarking	Selection of additional evaluation benchmarks
	Benchmark Creation	Development of novel benchmarks to evaluate identified concerns
Documentation	Warning	Documentation of general or use case-specific limitations
	Non-Use	Documentation of cases where the dataset should not be used at all
Packaging and Release	Licensing	Development of appropriate licensing and terms of use specifications
	Access	Development of limited access policies

Table 2: A non-exhaustive list of actions that may be taken to address social risks identified in data.

datasets available under a CC-BY 4.0 license. We apply the analyses from the standpoint of an individual or team seeking to repurpose data for their own use. In this vein, our (hypothetical) goal is not to generate scientifically novel results, but rather to develop derivative datasets from C4 and LAION-400M while assessing representational biases that have been previously identified in text and image datasets. We do not apply the framework in its entirety to each example; instead, we focus on a few key analyses specific to questions of representational harm. We do this for two reasons. First, not every analysis detailed in our framework is relevant for understanding a specific social group or modality. Thus, in practice, only a selection of analyses from the full framework will be conducted. Second, some analyses are not yet technically feasible and techniques to achieve them are themselves the subject of research (e.g., detecting hateful symbols and memes in images (Mathias et al. 2021)). Finally, because we focus on how to use results to inform risk mitigation, we do not detail the classifiers we use.

Analyzing Gender in C4

Gender bias is widely explored in language modeling, for example in translation (Stella 2021; Savoldi et al. 2021) and coreference resolution (Rudinger et al. 2018) tasks. At the same time, practitioners must still run analyses to understand known gender bias in their own data and make decisions about how to address biases that emerge. Taking the perspective of a product team seeking to proactively audit gender bias in the development of a new language dataset, we ask 1) what genders are represented in *our* data, and 2) what content is associated with different genders? To answer the first question, we look to the Human Factor and Human $\times$ Content analyses. We focus on gender pronouns and identity terms (i.e., "woman", "boy") as explicit gender references in text using a selection of gendered ICA webref entities. For simplicity, we calculate the top 20 TF-IDF tokens per document, filtering out tokens that are punctuation, numbers, and URL components. Next, to analyze content associations, we focus on the document topics most often associated with each gendered webref entity using a topic classifier similar to Google Cloud Natural Language API[4]. Due to the scope of this example, we do not calculate the average toxicity of documents in which gender terms appear. Across a dataset of this size, average toxicity would be more meaningful to consider alongside disaggregated calculations to determine how document sources differently contribute to toxicity.

Output: For the purposes of our example, we focus on an abbreviated set of results. Results can be seen in Figure 2. The distribution of binary gender pronouns shows some imbalance. C4 contains just under 50 million female pronouns, approximately 80 million male pronouns, and an unknown count of nonbinary pronouns. Our results lack nonbinary pronoun usage due to the lack of reliable method for disambiguating between plural and singular uses of "they" and "them". However, we anticipate extremely low representation of singular uses of "they", as well as other nonbinary pronouns such as "Xe" and "Ze" (Dev et al. 2021b). The distribution of webref entities shows a similar skew in references to women and girls compared with men and boys. Each entity can be interpreted as representing a range of terms related to the ones shown. Similarly, the top-associated topics with binary gender terms are similar, with entertainment media appearing more often for women. Average toxicity for each gender entity is generally low, with the highest average toxicity for "transsexual".

Actions: The collection of analyses provides a better understanding of how gender is depicted than any of the analyses alone, though it is critical to note that these results are limited to binary gender. Thus the calculated distributions are not reliable for decisions regarding nonbinary representation, however it is likely that nonbinary representation is limited and therefore worth considering in potential supplemental data collection and downstream model evaluation.

Potential negative effects of the binary pronoun imbalance may be lessened by the more balanced representation of gender terms, though the top topics for gendered terms suggest relatively more associations between women and celebrity topics such as beauty and gossip. For the development of a robust benchmark, for example, an Action to consider is

[4]https://cloud.google.com/natural-language/docs/classifying-text

to rebalance the dataset to create a more equal distribution of binary gender pronouns. This could be achieved by using techniques such as duplicating data and swapping pronouns, which can mitigate gender biases without significant effects on benchmark performance (Zhao et al. 2018). Such rebalancing should give attention not only to overall pronoun frequency but also topic frequency in which gendered terms occur, for example by resampling data from a subset of sources with more equal topic distributions. Alternatively, in a context with limited resources, mitigation should include documentation and flagging for downstream evaluation.

Evaluating Age Representation in C4

We motivate our analyses based on prior research that identifies age bias as an issue for ML and AI development (Díaz et al. 2018; Garcia de Alford et al. 2020) and calls for increased representation of older adults in AI datasets (Park et al. 2021). Given the underrepresentation of older adults in datasets and challenges to boosting representation in web-scraped data (Díaz et al. 2018), we assess the nature of older adult depictions in C4 using Association analyses with the goal of improving pre-training data mixtures with limited access to additional data. We assess age depictions using top-associated tokens and topics with age-related terms.

Output: We see in Figure 2 the tokens most associated with old age terms, which occur 110,000 times in the dataset. These include dementia and degeneration, both of which can render negative sentiment. We see related associations in the topic analysis. Topics include health topics, including medical conditions and assisted living, as well as skin and face care beauty products, which likely point to content covering anti-ageing products and discussion.

Action: In line with prior work, we find both low and skewed representation of older age in text. If a data collection or generation pipeline is feasible to pursue, targeted data collection or synthetic data could be used to rebalance the data, ideally through ablation studies to understand how model performance changes. Some work has been conducted on decoupling adjective associations from select identity terms (Dev et al. 2021a), however broader sentential context surrounding age-related terms may still carry negative or stigmatized sentiment. Filtering data has been used to mitigate age biases in prior work (Díaz et al. 2018), and may be appropriate for targeting specific terms, such as "frail", even though data representation is already low. Toward the goal of improving data mixtures, other actions may be taken. Analysis results can also be flagged to evaluate the downstream model for similar biases to assess model limitations. Contingent on these evaluations, documentation warnings or non-use cases may also be necessary.

Analyzing Queer Representation in LAION-400M

LAION-400M is a multimodal text and image dataset that features over 400 million image-text pairs extracted from Common Crawl, prominently used in text-to-image generation. The dataset is unstructured and uncurated, though it does feature NSFW tags, which were used to identify and filter a number of illicit images. Critically, a number of text-image datasets have been shown to reflect various representational biases; for example, DALL-E and its variants reproduce gender and skin tone biases from neutral prompts (Cho, Zala, and Bansal 2022). We apply the framework to LAION-400M to analyze such multimodal and text associations.

Researchers have found unintended and undesirable associations with queer identity terms in text datasets (Dixon et al. 2018) and sexually explicit depictions of women in LAION-400M (Birhane and Prabhu 2021; Birhane, Prabhu, and Kahembwe 2021). While explicit content may be useful to preserve for specific applications, unintentional inclusion risks accidental generation of explicit content by downstream models. Knowing that queer identity terms are often associated with sexually explicit topics in text, we assess whether these concerns appear across the combination of text and image modalities, again considering mitigations for adapting a third-party training dataset. We look to the results of Association analyses in text using the same topic classifier from our prior analyses, and we run multimodal Association analyses using a classifier similar to Google Cloud Vision API[5], which identifies sexual content in images. We analyze social identity terms via webref entities.

Output: The top-associated tokens with each queer identity term, which can be seen in Figure 3, largely reference other queer identity terms. Figure 3 shows the top topics associated with a variety of sexual identities. Prominent among these are those that seemingly refer to various sexual topics and activities. This includes for "heterosexuality". Interestingly, the most frequent topic for "heterosexuality" is "LGBT Porn" which suggests that the term is connected to a subgenre of pornographic videos. Though not a sexual identity, the generally derogatory term "transsexual" is also strongly associated with sexual terms and topics.

Action: Considering dataset usage for T2I training in particular, there is likely a very limited set of use cases in which the generation of sexual content would be appropriate. Such use cases would likely entail very specialized dataset curation and model development. Therefore, one could consider filtering sexual content in order to both limit its downstream production as well as to avoid the inclusion of published sexual content, which has often been made public without sex worker consent (Cole 2020). Because filtering may remove nearly half of the instances of some identity terms, rebalancing may also be needed. Alternatively, sexual content in text data could be augmented with tags to preserve a downstream model's ability to detect it while limiting its production.

Evaluating Representation in Data

In line with Mitchell et al. (2022)'s call to establish practices for measuring data, characterizing how people are represented in data is a necessary part of identifying downstream risks. Yet, RAI lacks systematized guidance to address the myriad ways that social identity emerges across data modalities. Up to this point, there has been little guidance for triangulating social identity in data using combinations of features and analyses to measure representation.

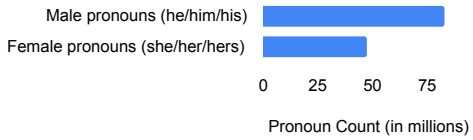
[5]https://cloud.google.com/vision/docs/reference/rest/v1/images/annotate#safesearchannotation

Dataset Analysis Output: C4

Pronoun Distribution:

Task: Distribution of gender pronouns that occur
Analysis Object: Text
Effort: Low
Dependencies: None
Action: Flag pronoun distribution

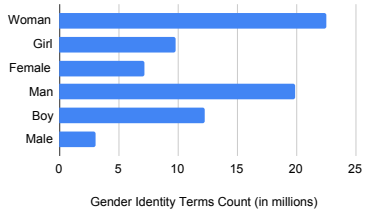
Output:



Gender Term Distribution:

Task: Calculate proportion of text referencing different social identity groups, considering unitary and intersectional groups
Analysis Object: Text
Effort: Low
Dependencies: None
Action: Flag social identity representation

Output:



Gender Terms × Topic:

Task: Calculate the distribution of topics, disaggregated by gender identity terms
Analysis Object: Inferred Signals + Text
Effort: Low
Dependencies: Topic Classifier
Action: Flag problematic associations

Output:

Woman	Books/Lit, History, Entertainment
Girl	Online Media, Journals, Entertainment
Female	Books/Lit, Entertainment, Online Media
Man	Books/Lit, History, Religion
Boy	Books/Lit, History, Religion
Male	Medical Lit, Books/Lit, Bio. Sciences

Age Terms × Top Tokens:

Analysis Object: Text
Task: For social identity terms in text, calculate top word co-occurrences for each.
Effort: Low
Dependencies: None
Action: Flag associations. Consider rebalancing harmful associations that emerge in model evaluations

Output:

Ageing	anti-aging, wrinkles, collagen, elasticity, age-related
Old Age	age-related, middle-aged, degeneration, dementia, ripe
Elderly	dementia, ageing, older, alzheimer, frail

Age Terms × Topic:

Task: Calculate the distribution of topics within text, disaggregated by social identity terms or inferred social identity signals
Analysis Object: Inferred Text Signals + Text
Effort: Low
Dependencies: Topic Classifier
Action: Flag problematic associations

Output:

Ageing	Medical Lit, Skin Care, Face Care, Anti-Ageing, Nutrition
Old Age	Medical Lit, Books/Lit, Health Cond., History, Retirement
Elderly	Medical Lit, Retirement, Assisted Living, Geriartics, Health Cond.

Figure 2: Sample dataset analyses output for C4.

Dataset Analysis Output: LAION-400M

Queer Identity Terms × Sexual Imagery:

Analysis Object: Text — Inferred Image Signals

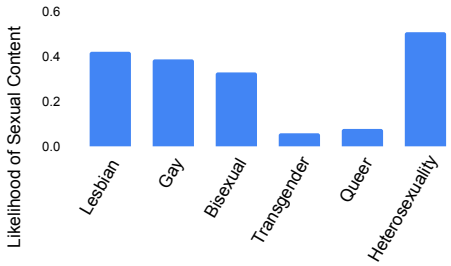
Task: Calculate co-occurrences between social identity terms and sexual imagery.

Effort: Low

Dependencies: Visual Content Classifier

Action: Data filtering or tagging. Possible rebalancing

Output:



Queer Identity Terms × Topic:

Task: Calculate the distribution of topics by queer identity terms

Output:

Analysis Object: Inferred Text Signals, Text

Effort: Low

Dependencies: Topic Classifier

Action: Follow up analyses of sexual content

Lesbian	Adult Videos, LGBT Porn, Porn
Gay	LGBT, LGBT Porn, Online Images
Bisexual	LGBT, Porn, Online Images
Transgender	LGBT, Social Issues, Discrimination
Queer	LGBT, Events/Listings, Online Images
Heterosexuality	Porn, Adult Videos, Porn

Figure 3: Sample dataset analyses output for LAION-400M Dataset.

Notably, our framework omits both a canonical list of social identities to analyze as well as an exhaustive list of methods to measure a given social identity. We do this because social identity is unstable nature (Hanna et al. 2020). At the same time, the features people readily associate with social identities changes with context, complicating algorithmic fairness evaluations, which have often failed to recognize the social construction of identities such as race (Benthall and Haynes 2019). Beyond specific data modalities, the characteristics associated with social identities change with context and, in a single context, the same features can be associated with more than one identity group. For example, Hispanic surnames are prevalent in both Latin America and the Philippines, meaning an analysis of cultural diversity cannot rely on these features alone. In other words, the salience of social cues in one context can change considerably in another, and previously meaningless signals can take on new significance. Scholars in FAccT have explored these concerns in cross-cultural explorations of algorithmic fairness, which includes cultural-specificity in the relevant axes along which discrimination occurs, as well as the meaning of fairness itself (Sambasivan et al. 2021a).

Analyzing how people are depicted is additionally challenging because "good" social representation can change with context. Chasalow and Levy's characterization of representativeness as a concept that is both time and place specific also applies more broadly to analyses of people in data (Chasalow and Levy 2021). The social categories we attend to are shaped both by normative assumptions about what should be measured as well as the existence of a name or conception of a social category. For example, (Andrews, Powell, and Ayers 2022)'s research suggests that a hypothetical word list generated today to analyze disability representation in a dataset would likely feature different terminology than a list generated 30 years ago. As new language and visual cues emerge that describe existing or newly salient social groups, our approaches to analyzing group representation must be updated. Shifts in terminology point to a need for RAI processes to be easily repeatable and updateable in order to support revisiting of canonical datasets, such as those used for benchmark evaluations, with updated attention to emergent or different social identity cues. In orienting our framework around high-level questions about how people are represented, we seek to make the analyses flexible to such shifts. Moreover, the acceptable thresholds of disparities between categories for the same analysis and in the same application context may also vary over time according to concerns for specific social risks as well as the development of relative social disparities in society more broadly.

The Role of Datasets in Assessing Harm

In developing a framework to help support these analyses and development decisions, we respond to and extend work that has advocated for more care and attention in data work (Sambasivan et al. 2021b), including the growing focus on data-centric AI (DCAI). Data-centric AI aims to reconfigure

AI development to "stretch [data's] lifespan beyond the so-called 'pre-processing' step" (Jarrahi, Memariani, and Guha 2022). Building from DCAI's focus on understanding the data used and produced throughout ML development, our framework sets a foundation for determining how dataset analyses of social identity features should be conducted. If DCAI aims to shift focus from the model to the data, our framework pushes on a human-centered angle within that data focus. Some work in DCAI does attend to data sources and the sociocultural views they represent, such as those expressed through data annotation (Díaz et al. 2022; Arhin et al. 2021; Mishra and Gorana 2021). However, this work focuses on human-labeled data, which is difficult to obtain at the scale foundation model training datasets.

Because we explicitly avoid mapping a specific distribution or association to fixed downstream risks, an important area for future work in DCAI and RAI development is to study the effects of different distributions and mixtures on model performance, both in terms of output bias and typical performance metrics such as accuracy. An important consideration for applying any dataset evaluations is determining when a certain data distribution becomes a problem. While there is some recent work studying the downstream effects of data composition on the model (Longpre et al. 2023b), these are limited by development costs. Thus, it is not always clear which distributions in a dataset are a cause of downstream harms. However some data skews, such as overwhelmingly sexual words and imagery associated with the label "Asian" are arguably egregious enough to warrant rebalancing or removal as a form of proactive risk mitigation. These knowledge gaps point to opportunities to use our framework to determine whether certain data mitigations in combination with model architectures might correct undesired model behavior. In this way the framework is a concrete aid in data benchmarking, which refers to "strategies to compare the quality of data in training and test sets across two consecutive iterations (Jarrahi, Memariani, and Guha 2022)." Across iterative development of the same model, as well as across developments of distinct models, the framework can act as a consistent measuring stick for consistently relating representations in data to model fairness.

Supporting RAI Workflows

Beyond algorithm design choices, RAI develpment requires dataset analyses that integrate into existing RAI workflows in ways that complement existing tools and frameworks. This includes structured guidance to repeatedly apply analyses of human depictions in data in order to mitigate downstream risks. Related RAI development frameworks include the internal auditing framework developed by Raji et al., which outlines an end-to-end set of auditing processes and transparency artifacts to help ensure that development processes meet institutional standards and requirements (Raji et al. 2020). In the framework, the Artifact Collection stage includes the development of Model Cards (Mitchell et al. 2019) and Datasheets (Gebru et al. 2021). The framework supports the development of transparency artifacts by standardizing results and flagging benchmark tests to prioritize. In this context, it stands as a structured auditing aid.

In addition, our dataset analysis framework can also be used to guide dataset modifications in other contexts, such as in the development of benchmarks. For benchmarks as well as datasets aimed to be field-wide standards, more comprehensive responses to analysis results would ideally be undertaken. There is a tendency for ML benchmarks to play a role of "model organisms" (Chasalow and Levy 2021; Denton et al. 2020)—research objects that are popular within fields to the point of overuse and that limit the claims of research conducted upon them (e.g., per (Chasalow and Levy 2021) these include fruit flies in biology, chess in AI, and Twitter in social media research)—heightening concerns of dataset bias. Indeed, fairness evaluations in ML already tends to rely on a small number of benchmark datasets (Fabris et al. 2022). Our framework supports the development of new benchmarks, including those not focused on fairness concerns. The framework acts as a robustness evaluation on social differences in data that might otherwise go ignored.

How and when to act on analysis results depends on the nature of the data being used as well as risk mitigation goals. That is, the same results may warrant different action, depending on context. In addition, the high cost of foundation model training limits opportunities to run comprehensive studies to identify which mitigation strategies improve fairness issues as well as their impacts model performance generally. For RAI, this means making mitigation decisions with limited information about specific impacts– a challenge exacerbated by data cascades, which compound to produce out-sized, negative outcomes (Sambasivan et al. 2021b). Yet, the range of potential downstream risks requires proactive decision making. While decision making at the dataset level is difficult when downstream applications are unclear, the framework provides valuable information about people represented in data, and can be used to guide downstream evaluations of fine-tuning data or model-generated data.

Conclusion

The open-ended nature of AI risks and harms continues to pose challenges to RAI practitioners seeking to identify and mitigate risks, at times with limited information about how downstream models will be fine-tuned or applied. Moreover, as new modalities and combinations thereof emerge, RAI practitioners must determine how social identity manifests in data and make pragmatic decisions regarding the most relevant data features to assess in relation to social risks. Even when it is known how to identify identity group in data, a definitive list of all possible signals for a single group (e.g., women) is impossible to formulate due to sociocultural variation across place and time. Finding the "right" degree of comprehensive coverage, then, is an ongoing RAI challenge.

Ultimately, dataset evaluations are just one component of RAI, and research is needed to better understand the efficacy of dataset mitigations and how they interact with model architecture. These experiments already face resource constraints, however our framework is a step towards scaffolding investigations to assess mitigation impacts. Building from critical dataset audits and concurrent work standardizing these efforts, we provide our framework as a systematic grounding for managing this task.

References

Ahmadi, S.; and Masoud, M. 2020. Towards machine translation for the Kurdish language. *arXiv preprint arXiv:2010.06041*.

Andrews, E. E.; Powell, R. M.; and Ayers, K. 2022. The evolution of disability language: Choosing terms to describe disability. *Disability and Health Journal*, 15(3): 101328.

Anil, R.; Dai, A. M.; Firat, O.; Johnson, M.; Lepikhin, D.; Passos, A.; Shakeri, S.; Taropa, E.; Bailey, P.; Chen, Z.; Chu, E.; Clark, J. H.; Shafey, L. E.; Huang, Y.; Meier-Hellstern, K.; Mishra, G.; Moreira, E.; Omernick, M.; Robinson, K.; Ruder, S.; Tay, Y.; Xiao, K.; Xu, Y.; Zhang, Y.; Abrego, G. H.; Ahn, J.; Austin, J.; Barham, P.; Botha, J.; Bradbury, J.; Brahma, S.; Brooks, K.; Catasta, M.; Cheng, Y.; Cherry, C.; Choquette-Choo, C. A.; Chowdhery, A.; Crepy, C.; Dave, S.; Dehghani, M.; Dev, S.; Devlin, J.; Díaz, M.; Du, N.; Dyer, E.; Feinberg, V.; Feng, F.; Fienber, V.; Freitag, M.; Garcia, X.; Gehrmann, S.; Gonzalez, L.; Gur-Ari, G.; Hand, S.; Hashemi, H.; Hou, L.; Howland, J.; Hu, A.; Hui, J.; Hurwitz, J.; Isard, M.; Ittycheriah, A.; Jagielski, M.; Jia, W.; Kenealy, K.; Krikun, M.; Kudugunta, S.; Lan, C.; Lee, K.; Lee, B.; Li, E.; Li, M.; Li, W.; Li, Y.; Li, J.; Lim, H.; Lin, H.; Liu, Z.; Liu, F.; Maggioni, M.; Mahendru, A.; Maynez, J.; Misra, V.; Moussalem, M.; Nado, Z.; Nham, J.; Ni, E.; Nystrom, A.; Parrish, A.; Pellat, M.; Polacek, M.; Polozov, A.; Pope, R.; Qiao, S.; Reif, E.; Richter, B.; Riley, P.; Ros, A. C.; Roy, A.; Saeta, B.; Samuel, R.; Shelby, R.; Slone, A.; Smilkov, D.; So, D. R.; Sohn, D.; Tokumine, S.; Valter, D.; Vasudevan, V.; Vodrahalli, K.; Wang, X.; Wang, P.; Wang, Z.; Wang, T.; Wieting, J.; Wu, Y.; Xu, K.; Xu, Y.; Xue, L.; Yin, P.; Yu, J.; Zhang, Q.; Zheng, S.; Zheng, C.; Zhou, W.; Zhou, D.; Petrov, S.; and Wu, Y. 2023. PaLM 2 Technical Report. arXiv:2305.10403.

Arhin, K.; Baldini, I.; Wei, D.; Ramamurthy, K. N.; and Singh, M. 2021. Ground-Truth, Whose Truth?–Examining the Challenges with Annotating Toxic Text Datasets. *arXiv preprint arXiv:2112.03529*.

Barocas, S.; Crawford, K.; Shapiro, A.; and Wallach, H. 2017. The Problem With Bias: Allocative Versus Representational Harms in Machine Learning. SIGCIS.

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

Benthall, S.; and Haynes, B. D. 2019. Racial categories in machine learning. In *Proceedings of the conference on fairness, accountability, and transparency*, 289–298.

Birhane, A.; and Prabhu, V. U. 2021. Large image datasets: A pyrrhic win for computer vision? In *2021 IEEE Winter Conference on Applications of Computer Vision (WACV)*, 1536–1546. IEEE.

Birhane, A.; Prabhu, V. U.; and Kahembwe, E. 2021. Multimodal datasets: misogyny, pornography, and malignant stereotypes. *arXiv preprint arXiv:2110.01963*.

Blodgett, S. L.; Barocas, S.; Daumé III, H.; and 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*, 5454–5476. Online: Association for Computational Linguistics.

Bommasani, R.; Hudson, D. A.; Adeli, E.; Altman, R.; Arora, S.; von Arx, S.; Bernstein, M. S.; Bohg, J.; Bosselut, A.; Brunskill, E.; et al. 2021. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*.

Bommasani, R.; Klyman, K.; Longpre, S.; Kapoor, S.; Maslej, N.; Xiong, B.; Zhang, D.; and Liang, P. 2023. The foundation model transparency index. *arXiv preprint arXiv:2310.12941*.

Buolamwini, J.; and Gebru, T. 2018. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*, 77–91. PMLR.

Carlini, N.; Tramer, F.; Wallace, E.; Jagielski, M.; Herbert-Voss, A.; Lee, K.; Roberts, A.; Brown, T. B.; Song, D.; Erlingsson, U.; et al. 2021. Extracting Training Data from Large Language Models. In *USENIX Security Symposium*, volume 6.

Chang, J.; and Custis, C. 2022. Understanding Implementation Challenges in Machine Learning Documentation. In *Equity and Access in Algorithms, Mechanisms, and Optimization*, 1–8.

Chasalow, K.; and Levy, K. 2021. Representativeness in statistics, politics, and machine learning. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 77–89.

Cho, J.; Zala, A.; and Bansal, M. 2022. Dall-eval: Probing the reasoning skills and social biases of text-to-image generative transformers. *arXiv preprint arXiv:2202.04053*.

Chowdhery, A.; Narang, S.; Devlin, J.; Bosma, M.; Mishra, G.; Roberts, A.; Barham, P.; Chung, H. W.; Sutton, C.; Gehrmann, S.; et al. 2022. Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*.

Cole, S. 2020. These Nudes Do Not Exist and I Don't Know Why This Startup Does Either. In *Vice*. Https://www.vice.com/en/article/qjdx7w/these-nudes-do-not-exist-and-i-dont-know-why-this-startup-does-either.

Denton, E. L.; Hanna, A.; Amironesei, R.; Smart, A.; Nicole, H.; and Scheuerman, M. K. 2020. Bringing the People Back In: Contesting Benchmark Machine Learning Datasets. *ArXiv*, abs/2007.07399.

Dev, S.; Li, T.; Phillips, J. M.; and Srikumar, V. 2021a. OSCaR: Orthogonal Subspace Correction and Rectification of Biases in Word Embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 5034–5050. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics.

Dev, S.; Monajatipoor, M.; Ovalle, A.; Subramonian, A.; Phillips, J.; and Chang, K.-W. 2021b. Harms of Gender Exclusivity and Challenges in Non-Binary Representation in Language Technologies. In Moens, M.-F.; Huang, X.; Specia, L.; and Yih, S. W.-t., eds., *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 1968–1994. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics.

Díaz, M.; Johnson, I.; Lazar, A.; Piper, A. M.; and Gergle, D. 2018. Addressing age-related bias in sentiment analysis. In *Proceedings of the 2018 chi conference on human factors in computing systems*, 1–14.

Díaz, M.; Kivlichan, I.; Rosen, R.; Baker, D.; Amironesei, R.; Prabhakaran, V.; and Denton, E. 2022. Crowdworksheets: Accounting for individual and collective identities underlying crowdsourced dataset annotation. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, 2342–2351.

Dixon, L.; Li, J.; Sorensen, J.; Thain, N.; and Vasserman, L. 2018. Measuring and mitigating unintended bias in text classification. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, 67–73.

Dodge, J.; Sap, M.; Marasović, A.; Agnew, W.; Ilharco, G.; Groeneveld, D.; Mitchell, M.; and Gardner, M. 2021. Documenting large webtext corpora: A case study on the colossal clean crawled corpus. *arXiv preprint arXiv:2104.08758*.

Elazar, Y.; Bhagia, A.; Magnusson, I.; Ravichander, A.; Schwenk, D.; Suhr, A.; Walsh, P.; Groeneveld, D.; Soldaini, L.; Singh, S.; et al. 2023. What's In My Big Data? *arXiv preprint arXiv:2310.20707*.

Fabris, A.; Messina, S.; Silvello, G.; and Susto, G. A. 2022. Tackling documentation debt: a survey on algorithmic fairness datasets. In *Equity and Access in Algorithms, Mechanisms, and Optimization*, 1–13.

Friedler, S. A.; Scheidegger, C.; Venkatasubramanian, S.; Choudhary, S.; Hamilton, E. P.; and Roth, D. 2019. A comparative study of fairness-enhancing interventions in machine learning. In *Proceedings of the conference on fairness, accountability, and transparency*, 329–338.

Garcia de Alford, A. S.; Hayden, S. K.; Wittlin, N.; and Atwood, A. 2020. Reducing Age Bias in Machine Learning: An Algorithmic Approach. *SMU Data Science Review*, 3(2): 11.

Garg, N.; Schiebinger, L.; Jurafsky, D.; and Zou, J. 2018. Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences*, 115(16): E3635–E3644.

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

Han, X.; Zhang, Z.; Ding, N.; Gu, Y.; Liu, X.; Huo, Y.; Qiu, J.; Yao, Y.; Zhang, A.; Zhang, L.; et al. 2021. Pre-trained models: Past, present and future. *AI Open*, 2: 225–250.

Hanna, A.; Denton, E.; Smart, A.; and Smith-Loud, J. 2020. Towards a critical race methodology in algorithmic fairness. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*, 501–512.

Heger, A. K.; Marquis, L. B.; Vorvoreanu, M.; Wallach, H.; and Wortman Vaughan, J. 2022. Understanding Machine Learning Practitioners' Data Documentation Perceptions, Needs, Challenges, and Desiderata. *Proceedings of the ACM on Human-Computer Interaction*, 6(CSCW2): 1–29.

Hendricks, L. A.; Burns, K.; Saenko, K.; Darrell, T.; and Rohrbach, A. 2018. Women also snowboard: Overcoming bias in captioning models. In *Proceedings of the European conference on computer vision (ECCV)*, 771–787.

Holstein, K.; Wortman Vaughan, J.; Daumé III, H.; Dudik, M.; and Wallach, H. 2019. Improving fairness in machine learning systems: What do industry practitioners need? In *Proceedings of the 2019 CHI conference on human factors in computing systems*, 1–16.

Huang, P.-S.; Zhang, H.; Jiang, R.; Stanforth, R.; Welbl, J.; Rae, J.; Maini, V.; Yogatama, D.; and Kohli, P. 2019. Reducing sentiment bias in language models via counterfactual evaluation. *arXiv preprint arXiv:1911.03064*.

Hutchinson, B.; Prabhakaran, V.; Denton, E.; Webster, K.; Zhong, Y.; and Denuyl, S. 2020. Social biases in NLP models as barriers for persons with disabilities. *arXiv preprint arXiv:2005.00813*.

Jarrahi, M. H.; Memariani, A.; and Guha, S. 2022. The Principles of Data-Centric AI (DCAI). *arXiv preprint arXiv:2211.14611*.

Johnson, J. M.; and Khoshgoftaar, T. M. 2019. Survey on deep learning with class imbalance. *Journal of Big Data*, 6(1): 1–54.

Kearns, M.; Neel, S.; Roth, A.; and Wu, Z. S. 2018. Preventing fairness gerrymandering: Auditing and learning for subgroup fairness. In *International conference on machine learning*, 2564–2572. PMLR.

Kemelmacher-Shlizerman, I.; Seitz, S. M.; Miller, D.; and Brossard, E. 2016. The megaface benchmark: 1 million faces for recognition at scale. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4873–4882.

Kleinberg, J.; Mullainathan, S.; and Raghavan, M. 2016. Inherent trade-offs in the fair determination of risk scores. *arXiv preprint arXiv:1609.05807*.

Kop, M. 2021. Eu artificial intelligence act: The european approach to ai. Stanford-Vienna Transatlantic Technology Law Forum, Transatlantic Antitrust ....

Lanitis, A.; Draganova, C.; and Christodoulou, C. 2004. Comparing different classifiers for automatic age estimation. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 34(1): 621–628.

Lima, G.; Grgić-Hlača, N.; Jeong, J. K.; and Cha, M. 2022. The Conflict Between Explainable and Accountable Decision-Making Algorithms. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, 2103–2113.

Longpre, S.; Mahari, R.; Chen, A.; Obeng-Marnu, N.; Sileo, D.; Brannon, W.; Muennighoff, N.; Khazam, N.; Kabbara, J.; Perisetla, K.; et al. 2023a. The data provenance initiative: A large scale audit of dataset licensing & attribution in ai. *arXiv preprint arXiv:2310.16787*.

Longpre, S.; Yauney, G.; Reif, E.; Lee, K.; Roberts, A.; Zoph, B.; Zhou, D.; Wei, J.; Robinson, K.; Mimno, D.; and Ippolito, D. 2023b. A Pretrainer's Guide to Training Data: Measuring the Effects of Data Age, Domain Coverage, Quality, Toxicity. arXiv:2305.13169.

Mathias, L.; Nie, S.; Davani, A. M.; Kiela, D.; Prabhakaran, V.; Vidgen, B.; and Waseem, Z. 2021. Findings of the WOAH 5 shared task on fine grained hateful memes detection. In *Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021)*, 201–206.

Mishra, A.; and Gorana, Y. 2021. Who Decides if AI is Fair? The Labels Problem in Algorithmic Auditing. *arXiv preprint arXiv:2111.08723*.

Mitchell, M.; Luccioni, A. S.; Lambert, N.; Gerchick, M.; McMillan-Major, A.; Ozoani, E.; Rajani, N.; Thrush, T.; Jernite, Y.; and Kiela, D. 2022. Measuring Data. *arXiv preprint arXiv:2212.05129*.

Mitchell, M.; Wu, S.; Zaldivar, A.; Barnes, P.; Vasserman, L.; Hutchinson, B.; Spitzer, E.; Raji, I. D.; and Gebru, T. 2019. Model Cards for Model Reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19, 220–229. New York, NY, USA: Association for Computing Machinery. ISBN 9781450361255.

Park, J. S.; Bernstein, M. S.; Brewer, R. N.; Kamar, E.; and Morris, M. R. 2021. Understanding the Representation and Representativeness of Age in AI Data Sets. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, 834–842.

Paullada, A.; Raji, I. D.; Bender, E. M.; Denton, E.; and Hanna, A. 2021. Data and its (dis)contents: A survey of dataset development and use in machine learning research. *Patterns*, 2(11): 100336.

Pushkarna, M.; Zaldivar, A.; and Kjartansson, O. 2022. Data Cards: Purposeful and Transparent Dataset Documentation for Responsible AI. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*.

Raffel, C.; Shazeer, N.; Roberts, A.; Lee, K.; Narang, S.; Matena, M.; Zhou, Y.; Li, W.; and Liu, P. J. 2020. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *Journal of Machine Learning Research*, 21(140): 1–67.

Raji, I. D.; Smart, A.; White, R. N.; Mitchell, M.; Gebru, T.; Hutchinson, B.; Smith-Loud, J.; Theron, D.; and Barnes, P. 2020. Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*, 33–44.

Rostamzadeh, N.; Mincu, D.; Roy, S.; Smart, A.; Wilcox, L.; Pushkarna, M.; Schrouff, J.; Amironesei, R.; Moorosi, N.; and Heller, K. 2022. Healthsheet: development of a transparency artifact for health datasets. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, 1943–1961.

Rudinger, R.; Naradowsky, J.; Leonard, B.; and Van Durme, B. 2018. Gender bias in coreference resolution. *arXiv preprint arXiv:1804.09301*.

Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Satheesh, S.; Ma, S.; Huang, Z.; Karpathy, A.; Khosla, A.; Bernstein, M. S.; Berg, A. C.; and Fei-Fei, L. 2014. ImageNet Large Scale Visual Recognition Challenge. *CoRR*, abs/1409.0575.

Saleiro, P.; Kuester, B.; Hinkson, L.; London, J.; Stevens, A.; Anisfeld, A.; Rodolfa, K. T.; and Ghani, R. 2018. Aequitas: A bias and fairness audit toolkit. *arXiv preprint arXiv:1811.05577*.

Sambasivan, N.; Arnesen, E.; Hutchinson, B.; Doshi, T.; and Prabhakaran, V. 2021a. Re-imagining algorithmic fairness in india and beyond. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, 315–328.

Sambasivan, N.; Kapania, S.; Highfill, H.; Akrong, D.; Paritosh, P.; and Aroyo, L. M. 2021b. "Everyone wants to do the model work, not the data work": Data Cascades in High-Stakes AI. In *proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 1–15.

Savoldi, B.; Gaido, M.; Bentivogli, L.; Negri, M.; and Turchi, M. 2021. Gender Bias in Machine Translation. *Transactions of the Association for Computational Linguistics*, 9: 845–874.

Schiff, D.; Rakova, B.; Ayesh, A.; Fanti, A.; and Lennon, M. 2020. Principles to practices for responsible AI: Closing the gap. *arXiv preprint arXiv:2006.04707*.

Schuhmann, C.; Vencu, R.; Beaumont, R.; Kaczmarczyk, R.; Mullis, C.; Katta, A.; Coombes, T.; Jitsev, J.; and Komatsuzaki, A. 2021. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. *arXiv preprint arXiv:2111.02114*.

Shankar, S.; Halpern, Y.; Breck, E.; Atwood, J.; Wilson, J.; and Sculley, D. 2017. No classification without representation: Assessing geodiversity issues in open data sets for the developing world. *arXiv preprint arXiv:1711.08536*.

Srinivasan, R.; Denton, E.; Famularo, J.; Rostamzadeh, N.; Diaz, F.; and Coleman, B. 2021. Artsheets for Art Datasets.

Stella, R. 2021. A Dataset for Studying Gender Bias in Translation.

Thiel, D. 2023. Identifying and Eliminating CSAM in Generative ML Training Data and Models. Technical report, Technical report, Stanford University, Palo Alto, CA, 2023. URL https://purl ....

Torralba, A.; Fergus, R.; and Freeman, W. T. 2008. 80 million tiny images: A large data set for nonparametric object and scene recognition. *IEEE transactions on pattern analysis and machine intelligence*, 30(11): 1958–1970.

Wagner, B.; Rozgonyi, K.; Sekwenz, M.-T.; Cobbe, J.; and Singh, J. 2020. Regulating transparency? Facebook, twitter and the German network enforcement act. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 261–271.

Weidinger, L.; Mellor, J.; Rauh, M.; Griffin, C.; Uesato, J.; Huang, P.-S.; Cheng, M.; Glaese, M.; Balle, B.; Kasirzadeh, A.; et al. 2021. Ethical and social risks of harm from language models. *arXiv preprint arXiv:2112.04359*.

Wilson, B.; Hoffman, J.; and Morgenstern, J. 2019. Predictive inequity in object detection. *arXiv preprint arXiv:1902.11097*.

Zhao, J.; Wang, T.; Yatskar, M.; Ordonez, V.; and Chang, K.-W. 2018. Gender bias in coreference resolution: Evaluation and debiasing methods. *arXiv preprint arXiv:1804.06876*.