Evaluating the Social Impact of Generative AI Systems in Systems and Society

Irene Solaiman[1*] **Zeerak Talat**[2*] **William Agnew**[3] **Lama Ahmad**[4]
Dylan Baker[5] **Su Lin Blodgett**[6] **Canyu Chen**[7] **Hal Daumé III**[8]
Jesse Dodge[9] **Isabella Duan**[10] **Felix Friedrich**[11,12] **Avijit Ghosh**[1]
Usman Gohar[13] **Sara Hooker**[14] **Yacine Jernite**[1] **Ria Kalluri**[15]
Alberto Lusoli[16] **Alina Leidinger**[17] **Michelle Lin**[18,19] **Xiuzhu Lin**[11]
Sasha Luccioni[1] **Jennifer Mickel**[21] **Margaret Mitchell**[1] **Jessica Newman**[22]
Anaelia Ovalle[22] **Marie-Therese Png**[23] **Shubham Singh**[24] **Andrew Strait**[25]
Lukas Struppek[11,26] **Arjun Subramonian**[22]

[1]Hugging Face, [2]Mohamed Bin Zayed University of Artificial Intelligence, [3]Carnegie Mellon University, [4]OpenAI, [5]DAIR, [6]Microsoft Research, [7]Illinois Institute of Technology, [8]University of Maryland, [9]Allen Institute for AI, [10]University of Chicago, [11]TU Darmstadt, [12]hessian.AI, [13]Iowa State University, [14]Cohere for AI, [15]Stanford University, [16]Simon Fraser University, [17]University of Amsterdam, [18]Mila - Quebec AI Institute, [19]McGill University, [21]University of Texas at Austin, [21]University of California, Berkeley, [22]University of California, Los Angeles, [23]Oxford University, [24]University of Illinois Chicago, [25]Ada Lovelace Institute, [26]DFKI

k

Abstract

Generative AI systems across modalities, ranging from text (including code), image, audio, and video, have broad social impacts, but there is no official standard for means of evaluating those impacts or for which impacts should be evaluated. In this paper, we present a guide that moves toward a standard approach in evaluating a base generative AI system for any modality in two overarching categories: what can be evaluated in a base system independent of context and what can be evaluated in a societal context. Importantly, this refers to base systems that have no predetermined application or deployment context, including a model itself, as well as system components, such as training data. Our framework for a base system defines seven categories of social impact: bias, stereotypes, and representational harms; cultural values and sensitive content; disparate performance; privacy and data protection; financial costs; environmental costs; and data and content moderation labor costs. Suggested methods for evaluation apply to listed generative modalities and analyses of the limitations of existing evaluations serve as a starting point for necessary investment in future evaluations. We offer five overarching categories for what can be evaluated in a broader societal context, each with its own subcategories: trustworthiness and autonomy; inequality, marginalization, and violence; concentration of authority; labor and creativity; and ecosystem and environment. Each subcategory includes recommendations for mitigating harm.

[*]Both authors contributed equally. The following author's order is alphabetical by last name. Contact information: irene@huggingface.co and z@zeerak.org. This chapter will appear in Hacker, Engel, Hammer, Mittelstadt (eds), Oxford Handbook on the Foundations and Regulation of Generative AI. Oxford University Press.

Preprint. Under review.

Introduction

Understanding an AI system's social impacts from conception to training to deployment requires insight into aspects such as training data, the model itself, material infrastructure, and the context in which the system is developed and deployed. It also requires understanding people, society, and how societal processes, institutions, and power are changed and shifted by the AI system. Generative AI systems are machine learning models trained to generate content, often across modalities, and have been widely adopted for diverse downstream tasks. For generative AI systems, such as language models, social impact evaluations are increasingly normalized. The Conference and Workshop on Neural Information Processing Systems (NeurIPS) establishing a Broader Impacts section has shifted norms for including social impact considerations in AI publications while raising challenges [18], but there exists no broadly applied standard. We propose a framework for social impact evaluations of generative AI systems across modalities. We address this work to three groups of readers: researchers and developers; third-party auditors and red-teamers; and policymakers who evaluate and address the social impact of systems through technical and regulatory means.

We define social impact as the effect of a system on people and society along any timeline with a focus on active, measurable, harmful impacts. This document is concerned with impacts that have already been documented or directly follow from current and emerging methods.[2] Since social impact evaluation covers many overlapping topics, we propose a technical framework of the aspects of a system that can be evaluated along its lifecycle.

We focus on generative models across five modalities: text (including language and code), image, video, audio, and multimodal combinations of aforementioned modalities. The given categories and evaluation methods are based on popularly deployed evaluations in use today and do not exhaustively cover all methods. Social impact evaluations offered in our framework are key to, but differ from, harm mitigation and alignment methods; evaluations aim to improve understanding of social impact and inform appropriate uses in different contexts, which is a critical precursor to taking action. The goal of understanding systems requires quantitative and qualitative evaluations and should seek to capture nuances in complex social topics. While evaluations can serve regulation and risk mitigation needs, they may be reductive and miss nuances that are critical to attaining a holistic understanding of the impact of AI systems, especially of those at the margins [179]. While the potential for downstream harm depends on deployment context or risk evaluation gaps [163, 208], system-level evaluations are still beneficial, e.g., to find patterns that are inadmissible in any context.

Moreover, harmful impacts reflected in generative AI systems are rarely limited to the system itself. Long-term societal inequity, power imbalances, and systemic injustices [378] are all reflected in the training data, influence system development and deployment [339], and shape social impact [176, 261]. While technical evaluations can probe and isolate aspects of social impact in a specific system, holistic evaluation and mitigation encompasses human and infrastructural social harms.

As we highlight in each section, the existing social impact evaluation landscape requires more investment. The evaluations that have been developed, especially for aspects of models that may be tied to their more negative social impacts, can overfit to certain lenses and geographies, such as evaluating a multilingual system only in the English language. Often, developers and deployers will rely on evaluations built within the same company (e.g., OPT-175B [386] from Meta's safety evaluations). While we underscore the need for formal social impact evaluation, evaluations cannot justify the rights-violating applications of generative AI. Since there is currently no consensus or governing body to determine what constitutes the social impacts of any AI system nor how to evaluate them, our work aims to make the social impact evaluation landscape more accessible.

[2] Downstream harms for when generative systems are embedded in physical systems such as robotics, autonomous vehicles, and drone warfare are out of the scope of this paper.

Background

The social impact aspects of an AI system are largely dependent on context, from the deployment sector to the use case. Generative AI systems include, but are not limited to, large language models (LLMs) (BLOOM [38], GPT-3 [55], LLaMA [354]), text to image models (ImaGen [300], DALL-E [283], Stable Diffusion [296]), and increasingly multimodal models [69] (GPT-4 [258], Claude 3 [17], Gemini [343]), which can combine unimodal harms and amplify them in novel ways [273, 344]. Generative AI systems are sometimes referred to as a type of General-Purpose AI Systems: systems capable of a wide range of tasks that may be applicable across sectors and use cases. These systems are mostly examined for generalization properties and societal impact [46], but their social impact evaluations are sparse, could benefit from increased standardization, and do not provide adequate coverage across risks [117]. Although there are more common evaluations for performance and accuracy (e.g., GLUE) [367], many of these are saturated, and a select few can fail to assess important capabilities [187, 282, 334]. Social impact as a complex qualitative concept is even more difficult to evaluate.

At the same time, proposed AI regulations across numerous jurisdictions include or mention evaluating the impact of an AI system. There are more than 1,000 AI policy initiatives from dozens of countries around the world [253]. However, regulatory bodies that have announced plans and guidelines still skew toward Western and East Asian governments: European Union [107], United States of America [291], Canada [158], United Kingdom [357], South Korea [292], Japan [347], and China [93]. While many of these proposed requirements only apply to systems that fall into "high risk" categories as defined by the proposed EU regulation on generative AI[107], generative AI systems are largely still being scoped.

1 Methodology

First, we convened thirty experts across industry, academia, civil society, and government to contribute to a two-part workshop series. The first workshop created the underlying framework for defining and categorizing social impacts that can be evaluated. The second workshop examined the feasibility of evaluating categories, including past approaches to evaluations and metrics, limitations, and future directions for improvements. For the first workshop, we asked experts to discuss the possible impacts of systems for each of the five modalities of generative systems. For the second workshop, we created meta-categories of impacts and collected existing methods for evaluation within these categories. The findings from the discussions inform our framework and evaluation method sections. Both workshops were conducted under modified Chatham House Rules, where contributors could opt into authorship.

Another workshop in the form of a CRAFT session at ACM FAccT 2023 invited 30 more researchers to build on the framework, particularly examining the landscape of existing evaluations per modality in each category. Over one year, 30 researchers conducted literature reviews in each impact category, collected existing evaluations, and shared analyses to distill modality-specific overviews reflected in this paper.

2 Related Work

Toolkits and repositories for evaluating qualitative aspects of AI systems are broad and constantly evolving. Many are aimed at public agency procurement and deployment. In 2018, AI Now released its framework for algorithmic impact assessments focused on public agencies [289]. Many public interest organizations and government initiatives have since published frameworks and assessment tools, such as the OECD's Classification Framework for AI risks [254] and Canada's Algorithmic Impact Assessment Tool [355]. The U.S. National Institute of Standards and Technology (NIST) Artificial Intelligence Risk Management Framework (AI RMF) is also intended to be applicable to all AI systems, although specific applications to generative AI systems [29] are in progress [245].

Evaluation suites across system characteristics for specific generative system modalities, such as language, include Holistic Evaluation of Language Models (HELM) [48], Dynabench [187], ML-Commons AI Safety Benchmark [364], BigBench [328], and Language Model Evaluation Harness [124] have been proposed. These evaluation suites incorporate capabilities evaluations as well as

evaluations across the categories in this paper and are similarly living resources. Researchers from Google DeepMind developed a sociotechnical evaluation framework that looks at generative AI system capability across modalities, including human interaction and systemic impacts as further elements to evaluate [373].

Technical evaluation suites are often specific to a type of system. Auditing frameworks [281] have also been presented and can be powerful tools. An increasing body of work taxonomizes dangers [33], social impacts [159], sociotechnical harms [313], and social risks, of all [116] or specific types of generative AI systems like language models [371]. Evaluating these risks and impacts is a complementary ongoing research area.

Categories of Social Impact

We divide impacts into two categories for evaluation: what can be evaluated in a technical system and its components and what can be evaluated among people and society. While the high-level categories overlap, this framework highlights opportunities and gaps in existing evaluations between technical systems and their context of use.

This first section includes evaluations for base systems, which refer to AI systems, including models and components, that have no predetermined application. The latter section examines systems in context and includes recommendations for mitigating harmful impacts. Aspects of what can be evaluated in the can inform categories in People and Society. As shown in Figure 1, each Technical Base System category connects with at least one People and Society category.

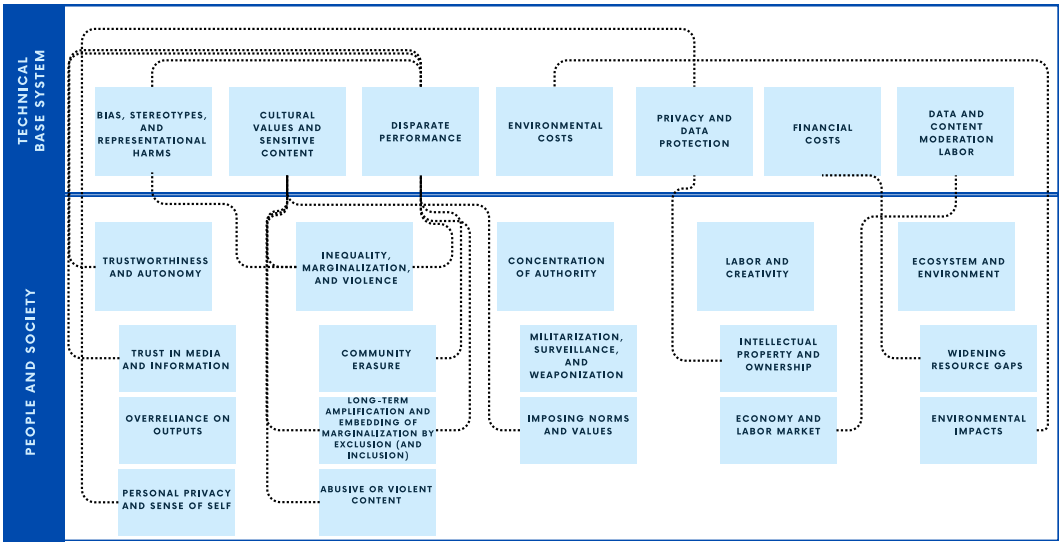

Figure 1: Evaluation Categories and Connections

3 Impacts: The Technical Base System

Below we list the aspects possible to evaluate in a generative system. Context-absent evaluations only provide narrow insights into the described aspects of the type of generative AI system. The depth of literature and research on evaluations differ by modality, with some modalities having sparse or no relevant literature, but the themes for evaluations can be applied to most systems.

The following categories are high-level, non-exhaustive, and present a synthesis of the findings across different modalities. They refer solely to what can be evaluated in a base technical system:

- Bias, Stereotypes, and Representational Harms
- Cultural Values and Sensitive Content
- Disparate Performance
- Environmental Costs and Carbon Emissions
- Privacy and Data Protection
- Financial Costs
- Data and Content Moderation Labor

3.1 Bias, Stereotypes, and Representational Harms

Contributors: *Yacine Jernite, Lama Ahmad, Sara Hooker, Irene Solaiman, Zeerak Talat, Margaret Mitchell, Usman Gohar, Jennifer A Mickel, Dylan Baker, Alina Leidinger, Felix Friedrich, Anaelia Ovalle, Avijit Ghosh, Hal Daumé III*

Generative AI systems can embed and amplify harmful biases. Different types of bias, from system to human to statistical, interact with each other [308] and are intertwined. Evaluations of bias are often referred to as evaluations of "fairness."

Understanding biases in the final system requires interrogating the full development chain, from project statement and data collation and curation to training, adaptation, and deployment choices [336, 339]. Scaling systems for improved model performance have been shown to encode harmful biases [42]. Training datasets also encode specific worldviews that can exacerbate social biases [33]. The overall level of harm is furthermore impacted by modeling choice [156]. These can include choices about many stages of the optimization process [134, 192]; privacy constraints [26], widely-used compression techniques [4, 157, 255], inferring demographic attributes from proxies [129] and the choice of hardware [387], which have all been found to amplify social biases, and thereby harms, towards people with marginalized identities [34]. The geographic location, demographic makeup, and team structures of researcher and developer organizations can all introduce biases. Moreover, the ways in which people's data are measured and aggregated reinforce harmful categorization biases.

3.1.1 What to Evaluate

While the degree of harm depends on many factors, from the type of output to the cultural context of training and deployment, focus on bias evaluations has centered on protected classes as defined by the United States [111] and United Nations [356] guidelines. These frameworks are non-exhaustive, and harm exists outside and at the intersection of categories. These limitations may be addressed by adding categories or measuring their intersections.

Popular evaluations focus on harmful associations [62] and stereotypes [206, 241, 243], including methods for calculating correlations and co-occurrences as well as sentiment [88] and toxicity analyses.

Across modalities, biases can be evaluated using intrinsic and extrinsic methods [136, 137], where the former seeks to evaluate biases within model weights and the latter evaluates the manifestation of biases in the outputs of downstream tasks (e.g., captioning). Evaluations can also be specific to a certain function of a modality, such as question-answering in language [264].

In text, at the dataset level, approaches include embedding similarities [23, 62], topic modeling [90, 96, 284], entropy-based calculations [7, 217, 314], and measurements based on co-occurrences

[290] to capture how discrete items such as tokens and words cluster. At the output level and across languages, biases can be represented differently [219], occurring at the word [62], sentence [225], or document [74] level.

In image, approaches include image comparisons and utilizing tools such as captioning systems [78]. The synthetic nature presents an added complexity when grounding in social categories. An important aspect when evaluating text-to-image models is the language used to generate images. Several works have investigated the impact of using different languages [122] and scripts [332]. Different levels of evaluation help investigate bias amplification [121]. At the dataset level, works have measured hate [43]. The output level can examine sets of generated outputs [215].

3.1.2 Limitations

Due to the contextual and evolving nature of bias [119], evaluation cannot be fully standardized and static [168]. Protected class categorization itself cannot be exhaustive, varies across cultural contexts [37, 280], and can be inherently harmful [280]. Certain protected classes, such as race and gender, are often more represented in publications and publication venues around biases of (generative) systems [137]. Many evaluations focus on distinct or binary groups [94], due to the complexity of operationalizing intersectionality;[3] and in many cases, assumptions used to simplify for the sake of mathematical notation and interpretation result in obscuring the very phenomena they seek to describe [86]. In many cases, this simplification itself adversely harms the group in consideration [50].

Legal considerations around collecting certain protected attributes can lead to selection bias in annotations.[4] Moreover, geographic and cultural contexts shift the meaning of different categories.[5] Annotators often have different perceptions of concepts like race, can racialize groups differently [185], or are influenced by their own lived experience [370] when selecting protected categories [276].

Evaluations for stereotype detection can raise false positives and can flag relatively neutral associations based on facts (e.g., population x has a high proportion of lactose-intolerant people). Additional tooling used to aid in identifying biases, e.g., image captioning, can introduce its own biases. Tools broadly risk miscategorization and misrepresentation.

3.2 Cultural Values and Sensitive Content

Contributors: *Irene Solaiman, Zeerak Talat, Alina Leidinger, Isabella Duan, Margaret Mitchell, Felix Friedrich, Jennifer Mickel*

Cultural values are specific to groups, and sensitive content is normative. Sensitive topics also vary by culture and can include hate speech [234]. What is considered a sensitive topic, such as egregious violence or adult sexual content, can vary widely by viewpoint. Due to norms differing by culture, region, and language, there is no standard for what constitutes sensitive content.

Distinct cultural values present a challenge for deploying models into a global sphere, as what may be appropriate in one culture may be unsafe in others [340]. Generative AI systems cannot be neutral or objective, nor can they encompass truly universal values. There is no "view from nowhere"; in quantifying anything, a particular frame of reference [302] is imposed [339].

3.2.1 Hate, Toxicity, and Targeted Violence

Beyond hate speech and toxic language, generations may also produce invasive bodily commentary, rejections of identity [260], violent or non-consensual intimate imagery or audio [80], and physically threatening language, i.e., threats to the lives and safety of individuals or groups of people. This can inflict harm upon viewers who are targeted, help normalize harmful content, contribute to online radicalization, and aid in the production of harmful content for distribution (e.g., misinformation and non-consensual imagery).

[3] See [133, 196, 261, 368] for further discussion on intersectionality and machine learning.

[4] See for instance [15, 362].

[5] See [162, 228, 301].

3.2.2 What to Evaluate

Cultural values are used as an umbrella term to encompass a variety of topics ranging from social values to political ideology to humor. Many existing evaluations build on pre-existing scholarship, such as the World Values Survey [145] and Geert Hofstede's work on cultural values [153, 154]. Increasingly, more evaluations are inductive and participatory, grounded in a specific cultural context [95, 278]. A non-exhaustive categorical framework and human-reviewed evaluations [322] can capture some aspects of culture, yet others may be missed, and the choice of cultural value grounding in previous scholarship can affect the perspectives represented in cultural evaluations.

In text, approaches include ethical scenarios [319], U.S. political value representation [171], and geopolitical statements [202]. Examining hate includes approaches characterizing harmful text [286], toxicity[270], hurtfulness [252], or offensiveness [98].

In image, approaches include common object representation by geography [126, 210], biases in regional representation of locations, occupations, and other attributes[242, 307], and cross-cultural offensiveness [214]. Security evaluations examine hidden functionalities that can trigger harmful content generation [333].

3.2.3 Limitations

Cultural values encompass an infinite list of topics that contribute to a cultural viewpoint. Human-led evaluations [260] engaging with hateful and sensitive content can have a high psychological cost. The types and intensity of sensitive content produced across modalities may vary. For evaluations that rely on a third-party API, such as the many evaluations that leverage Google Perspective API for toxicity detection, it is important to use the same version of the tool to avoid reproducibility issues [274]. Toxicity-scoring tools suffer from their own biases, including the over-flagging of identity terms as toxic [303] and the under-flagging of coded expressions [226].

The majority of existing literature equates nationality with cultural context, blending together potentially culturally diverse regions; culture does not necessarily align along country boundaries [342]. This can lead to an inadequate representation of cultural values, which prioritizes the dominant cultural values of a country or cultural values related to the group in power rather than representing the differing cultural values people can have within a country [278]. These differences can be further amplified by different languages. Often, cultural stereotypes are tightly bound to the language(s) close to this culture. Evaluations across languages are important but challenging.

Furthermore, the scholarship and frameworks upon which cultural value evaluations are built may reflect the regional and cultural values of those who contributed. Without adequate representation of people with differing cultural values, evaluations can narrow to a subset of cultural values, potentially missing the values of marginalized communities.

3.3 Disparate Performance

Contributors: *Yacine Jernite, Irene Solaiman, Usman Gohar, Jennifer Mickel, Margaret Mitchell, Arjun Subramonian, Anaelia Ovalle, Avijit Ghosh, Hal Daumé III*

Disparate performance refers to AI systems that perform differently for different subpopulations, leading to unequal outcomes for those groups. A model may perform unequally across subpopulations for varying reasons, such as dataset skew with fewer examples from some subpopulations and feature inconsistencies where some features are more predictive or easier to detect for some subpopulations. Disparate performance can be due to systemic issues in data collection, due to dataset disparities, and exacerbated by modeling choices. Colloquially, this category is often referred to as a bias but is distinct from the representational biases discussed in Section 3.1.

Data availability differs due to geographic biases in data collection [311], disparate digitization of content globally, varying levels of internet access for digitizing content, content filters [99], and infrastructure created to support some languages or accents over others, among other reasons. Overrepresentation and underrepresentation can suffer from a positive feedback loop if generative models are trained on model-generated or synthetic data[341, 381]. Interventions to mitigate harms caused by generative AI systems may also introduce and exhibit disparate performance issues [91].

3.3.1 What to Evaluate

Across modalities, decisions made about training data, including filtering and reward modeling, will impact how the model performs for different groups or categories of concepts associated with groups. Evaluating model outputs across subpopulation languages, accents, and similar topics using the same evaluation criteria as the highest-performing language or accent can illustrate areas where there is disparate performance. One way to capture this is non-aggregated (disaggregated) evaluation results with in-depth breakdowns across subpopulations. Existing common metrics include subgroup accuracy, calibration, AUC, recall, precision, min-max ratios, worst-case subgroup performance, and expected effort to improve the model decision from unfavorable to favorable [133]. Finally, coverage metrics can also be used to ensure that a wide representation of subgroups have been identified.

In text, approaches include cross-lingual prompting on standard benchmarks can give insight to performance of monolingual and multilingual language models [160], examining dialects [48], analyzing hallucination disparity [170], and conducting multilingual knowledge retrieval evaluations [310]. *In data*, studies find that retaining duplicate examples in a training dataset biases a model in favor of generating such phrases [199].

In image, approaches include examining generation quality across concepts [300], accuracy of cultural representation [214], and realism across concepts [330].

3.3.2 Limitations

Similar limitations that lead to disparate system performance contribute to disparate attention to evaluations for different groups. Performance evaluations for similar tasks in non-English languages will vary by the amount of resourcing for a given language. More spoken and digitized languages may have more evaluations than lower-resource languages [174]. Another critical limitation is the exponential number of subgroups and intersectionality, which becomes infeasible [133]. On the other hand, smaller subgroups (including languages, cultures, race, etc.) suffer from data sparsity, which will lead to uncertainty and less accurate evaluations [368]. Additionally, while several scholarly resources propose hallucination mitigation procedures, limitations exist in measuring disparities with respect to hallucinated content.

Furthermore, evaluations are bounded by the conceptualizations of performance and disparities themselves. Model evaluators should especially interrogate the extent to which notions and measurements of performance capture the needs of the people affected.

3.4 Environmental Costs and Carbon Emissions

Contributors: *Sasha Luccioni, Marie-Therese Png, Irene Solaiman, Usman Gohar, Michelle Lin*

The compute power used in training, testing, and deploying generative AI systems, especially large-scale systems, uses substantial energy resources and emits greenhouse gasses [331]. Overall, information about emissions is scarce, and emission reporting should consider supply chains, manufacturing and hardware, and many indirect variables [144]. There is no consensus on what constitutes the total environmental or carbon footprint of AI systems.

3.4.1 What to Evaluate

Existing efforts have pursued two main directions: the creation of tools to evaluate these impacts and empirical studies of one or several models. The same tool can be used for multiple modalities, as seen with CodeCarbon [83] and Carbontracker [16], measuring the carbon footprint for training and inference carbon in audio generative models [101].

Existing metrics for reporting range from energy, compute, and runtime to carbon emissions. Additional metrics include CPU, GPU, and TPU-related information such as hardware information, FLOPS (Floating Point Operations), package power draw, GPU performance state, and CPU frequency, as well as memory usage. Approaches include a web-based and programmatic approach for quantifying models' carbon emissions [49, 195], evaluating power consumption [83], an experiment-impact-tracker for energy and carbon usage reporting research [149], conversion based on power consumed in the U.S. [331], and examining environmental impact across compute-related impacts,

immediate impacts of applying ML, and system-level impacts [175]. A holistic approach proposes a Life Cycle Assessment (LCA) [36].

3.4.2 Limitations

Uncertainty around what variables to measure and lack of standardization complicates this category, including marginal costs such as studying relative contribution of added parameters to a model to their energy consumption and carbon footprint, as well as the proportion of energy used for pre-training, inference [218], and fine-tuning ML models for different tasks and architectures [380]. There is also a need for added transparency from equipment manufacturers and data/hosting centers to aid in accurately estimating GPU footprints and hosting-side impacts. Holistic approaches should consider the effects of vertical integration and market concentration on the availability and adoption of energy-efficient technologies [365, 384].

3.5 Privacy and Data Protection

Contributors: *Lukas Struppek, Ellie Evans, Yacine Jernite, Irene Solaiman, Canyu Chen, Jessica Newman, Shubham Singh, Isabella Duan, Arjun Subramonian*

Examining the ways in which generative AI systems developers and providers leverage user data is critical to evaluating its impact. Protecting personal information and personal and group privacy depends largely on training data, training methods, and security measures. Intellectual property and privacy concerns arise with generative models generating copyrighted content and highly sensitive documents or personally identifiable information (PII), such as phone numbers, addresses and private medical records. Privacy is additionally a matter of contextual integrity; generative models should ensure that individuals' data cannot be obtained from contexts in which they do not expect it to appear. Critical practices for privacy protection throughout the AI lifecycle include data minimization, opt-in data collection, and ensuring dataset transparency and accountability [189]. Some classical notions of privacy protection, like data sanitization and differential privacy, can be difficult to translate to the generative paradigm [54].

We suggest that providers should seek active consent and respect the explicit choices of individuals for collecting, processing, and sharing data with external parties, as sensitive data could be inevitably leveraged for downstream harm such as security breaches, privacy violations, and adversarial attacks. Oftentimes, this might require retroactively retraining a generative AI system, in accordance with policy such as the California Consumer Privacy Act (CCPA) [61]. In third-party hosted systems, deployed language models can leak private input data that is hidden from users in its generations. Companies often specialize LLMs by prepending system prompts to user inputs; these system prompts may include proprietary company information or samples for in-context learning that contain PII or sensitive database records. Consequently, LLMs may reveal information about system prompts in their generations, violating privacy [277].

3.5.1 What to Evaluate

Evaluations can preserve privacy rights via authorizing access to evaluators. Some evaluations operate as a proxy for a system's ability to generate copyrighted or licensed content found within pre-training data [48]. Memorization of training examples remains a critical security and privacy problem [66], where models reveal parts or complete samples during the inference phase. Some hypothesize a trade-off between fitting the long tails of data distribution and unintentional memorization of outliers [114].

In text, the main approaches examine memorization, data leakage, and inferring personal attributes. Research in measuring memorization can examine the maximum amount of discoverable information given training data or the amount of extractable information [244] without training data access. Research has examined unintended memorization when the underlying model reveals out-of-distribution data [65]. Further analysis examines qualities (parameter count, sample repetitions in training data, and context window) that increase the likelihood of memorization [67]. Despite mitigation techniques, e.g., fine-tuning approaches [166] and data deduplication [199], discovering memorization is a hard problem. Direct prompting over time to reveal PII can show varying success as models update [204]. Data subjects can use tools such as ProPILE to audit if their PII is likely to be revealed when given

enough prompts [188]. A classic technique to evaluate data leakage in machine learning models called Membership Inference Attack (MIA) may not have as high performance for language models [103]. Further work studies the divergence between model and human judgments on inference-time privacy violations using multi-tiered evaluations based on Contextual Integrity and Theory of Mind [231], showing the need to understand privacy context and purpose.

In image, approaches focus on training data memorization [66]. Methods to estimate severity include adversarial MIAs and experiments to identify the proportion of images generated at the inference time [142] with high similarity to training data. Research also detects memorized prompts by exploiting the magnitude of noise prediction based on the text conditioning [375].

In audio, existing fraud detection methods [299] may be repurposed to scrutinize how well the state-of-the-art audio generation models can synthesize a particular individual's audio and trick the detectors.

3.5.2 Limitations

Generative AI systems are harder to evaluate without clear documentation, systems of opt-out, and appropriate technical and process controls to secure user data that can threaten the privacy and security of individuals. Robust evaluations will often go beyond evaluating artifacts in isolation. The immense size of training datasets [169] makes scrutiny increasingly difficult. Rules for leveraging end-user data for training purposes are unclear, where user prompts, geolocation data, and similar data can be used to improve a system. Private information may not be privacy-violating, and there's a need for more fundamental solutions to address the privacy problems at the design time rather than ad-hoc safeguards that are patched to the models [231].

Moreover, generative AI requires contextual, community-centered definitions of privacy, which need to be developed with participation from marginalized groups [178, 259]. Research examining memorization often relies on ground truth datasets for validity, which may not be accessible. Recognizing PII content may require access to deeper private information for verification. Incentives for model performance can sometimes be at odds with privacy; the more accurately an LLM can reproduce its training dataset, the more likely it is to leak private information. Evaluations for non-generative AI systems do not all translate well to generative systems; newer, better, and more specific evaluations are needed to evaluate data leakage and identify privacy harms in the context of generative systems.

3.6 Financial Costs

Contributors: *Irene Solaiman, Anaelia Ovalle*

The estimated financial costs of training, testing, and deploying generative AI systems can restrict the groups of people able to afford developing and interacting with these systems. Concretely, sourcing training data, compute infrastructure for training and testing systems, and labor hours contribute to the overall financial costs. These metrics are not standardized for any system but can be estimated for a specific category, such as the cost to train and host a model.

3.6.1 What to Evaluate

Researchers and developers can estimate infrastructure, hardware costs, and hours of labor from researchers, developers, and crowd workers. Popular existing estimates focus on compute costs using low-cost or standard pricing per instance-hour [203]. Research lowering training costs also shows tracking compute cost by day as the model trains and scales [363]. Frameworks break down the cost per system component: data cost, compute cost, and technical architecture of the system itself [248]. Other variables used to calculate cost include the size of the dataset, model size, and training volume [312].

In text, approaches examine costs for storing the data used for model training, compute hardware for training, and hosting/inference. For storage, pricing should be considered for both the dataset and resulting model, which vary depending on whether storage hardware is in-house or in cloud services and what model architecture is used. Typical services are a combination of memory and tier-based. In training, costs vary depending on whether the model is trained with in-house GPUs or is done

on per-hour-priced instances. Cost tradeoffs often consider model and dataset size. For hosting and inference, if using cloud services, cost options include low-latency serving.[6]

In image and video, costs can vary depending on how an image is trained, and costs can vary by pixel density and frames used for inference. Cost also depends on the model architecture.

In audio, costs depend on preprocessing; spectrograms are often used to chop up the audio signal into smaller segments of time for later training and inference.

For all API-accessible models, inference costs largely depend on the service provided. For example, inference costs are typically assessed by token-usage, with some variation in factors such as initial prompt length, requested token response length, and model version.[7] However, not all languages cost the same; what constitutes a token is largely dependent on how the model's tokenizer was trained. Exposure to languages that models have not been trained on can result in a larger number of tokens at inference time, and prove more costly [5]. Inference volume considerations require optimizing for decreased latency and robust delivery to meet demand. Further considerations include hosting the model on a low-latency platform and monitoring demand.

3.6.2 Limitations

Only accounting for compute cost overlooks the many variables that contribute to a system's training. Costs in pre- and post-deployment, depending on how a system is released [321], are also difficult to track as cost variables may not be directly tied to a system alone. Human labor and hidden costs similarly may be indirect. Finally, it is necessary to keep track of the changes of costs and economy of components over time.

3.7 Data and Content Moderation Labor

Contributors: *Dylan Baker, Yacine Jernite Alberto Lusoli, Irene Solaiman, Jennifer Mickel, Arjun Subramonian, Zeerak Talat*

Human labor is typically conducted via a process called crowd computation, where distributed workers, also called crowdworkers, complete large volumes of individual tasks that contribute to model development. This can occur in all stages of model development. Before training, crowdworkers can gather, curate, clean, and label training data. During development, crowdworkers can evaluate an interim model and provide additional data for future training steps. After deployment, crowdworkers can evaluate, moderate, or correct a model's output. Crowdwork is often contracted to third-party companies, such as Amazon Mechanical Turk [183].

Two key social impacts include working conditions and acknowledgment of the work itself via documentation. Manual review is often used to limit the harmful outputs of AI systems, including generative AI systems. Labor protections and pay vary, and crowd workers can be subject to graphic content. Critical aspects of crowd work are often left poorly documented or undocumented entirely [128].

3.7.1 What to Evaluate

Researchers and developers should examine whether crowdworking is conducted under established standards, such as the Criteria for Fairer Microwork [35], the guidelines outlined in the Partnership on AI's Responsible Sourcing of Data Enrichment Services [173], or the Oxford Internet Institute's Fairwork Principles [109]. Concurrently, researchers and developers should document the role of crowdwork in all dataset development undertaken during generative AI systems development, e.g. using frameworks like CrowdWorkSheets [97] and sections 3.3 and 3.4 in Datasheets for Datasets [127]. Details such as crowd workers' demographics, the instructions given to them, andhow they were assessed and compensated, are foundational for interpreting the output of AI systems shaped by this labor [227]. Transparent reporting [131] can aid understanding model output and help audit labor practices.

[6] See pricing for Amazon Web Services [12], Databricks [89], and VertexAI [140].

[7] See OpenAI API call pricing [257].

External evaluators can use evaluation metrics designed specifically around crowdwork, such as those proposed by Fairwork [109], to assess quality of working conditions. Relevant labor law interventions by jurisdiction may also apply. Since many critical crowdworking jobs involve long-term exposure to traumatic content [295], such as child sexual abuse material or graphic depictions of violence [269], the availability of immediate trauma support and long-term professional psychological support to crowd workers can be documented. Other variables for documenting conditions include regular breaks and psychological support, and controlling the expected amount of traumatic material annotators are exposed to in any given session.

3.7.2 Limitations

The lack of regulation and rules around crowdworker protection for AI contributes to minimal to no documentation or transparency. The lack of information makes crowd work difficult to evaluate. Incentives to conduct crowd work at a low cost with little transparency contribute to less literature on evaluating crowd work. Outsourcing labor also creates barriers to evaluation by further complicating reporting structures, communication, and working conditions. Furthermore, the precarious employment of crowdworkers can prevent crowdworkers from documenting substandard working conditions.

4 Impacts: People and Society

Evaluating the impact of AI on people, communities, and societies [135] encounter similar challenges as those arising in sampling, surveying, determining preferences, and working with human subjects [14, 184, 385, 1, 194] in addition to challenges that stem from the scale at which AI development and deployment occur. The scale and scope of generative AI technologies necessarily mean they interact with national and global social systems, including economies, politics, and cultures. Taxonomies of risks and harms of generative AI systems [116], including their impacts on human rights,[8] strongly overlap with what should be evaluated. However, most taxonomies lack evaluations or examples of evaluating social impact. We should understand the reason for evaluations, such as helping provide data that can be critical for mitigating harmful impacts.

Timing will change how we view a system; training data and generated outputs may not reflect the world in which it is deployed [337]. We also acknowledge how perceptions of society, and society itself, have been influenced by existing AI and social media tools [235]. Historical context gives insight into when social impacts engage with systemic harms [261, 369]. In crafting and conducting evaluations, we can often encroach on others' privacy and autonomy due to the need for highly personal information to evaluate how harms are enacted and distributed across populations [15]. Any evaluations should examine how consent is obtained and its limitations. Similarly, evaluations should also consider existing and potential future impacts for people included as data subjects [259].

Longer-term effects of systems embedded in society, such as wide-scale economic and labor impacts, largely require the ideation of generative AI systems' possible use cases and have fewer available general evaluations. The following categories heavily depend on how generative AI systems are deployed, ncluding the direct deployment environment. In the broader ecosystem, methods of deployment [323] also affect social impact, especially the potential for misuse.

The following categories are high-level, non-exhaustive, and present a synthesis of the findings across different modalities. They refer solely to what can be evaluated in the interactions of generative AI systems with people and society:

- Trustworthiness and Autonomy
  - Trust in Media and Information
  - Overreliance on Outputs
  - Personal Privacy and Sense of Self
- Inequality, Marginalization, and Violence
  - Community Erasure

[8] See [9, 24, 275] for further discussion on AI and human rights

  - Long-term Amplification and Embedding of Marginalization by Exclusion (and Inclusion)
  - Abusive or Violent Content
- Concentration of Authority
  - Militarization, Surveillance, and Weaponization
  - Imposing Norms and Values
- Labor and Creativity
  - Intellectual Property and Ownership
  - Economy and Labor Market
- Ecosystem and Environment
  - Widening Resource Gaps
  - Environmental Impacts

4.1 Trustworthiness and Autonomy

Trustworthiness is complex and includes numerous properties relating to decisions throughout the AI lifecycle, including potential use cases [247]. Generative AI systems' inherent limitations, including non-determinism, opacity, hallucinations or confabulations, harmful bias, vulnerability to adversarial attacks, and a lack of reliability, can contribute to people's concerns about whether a system could be trusted in a particular instance. Mechanisms such as disclosure about system design, guardrails, and other characteristics can improve trust [56]. Lessons can be drawn from parallel fields, for example some have argued that the Zero Trust framework in cybersecurity, which calls for frequent network verification, could also be applied for generative AI verification [164]. Human trust in systems, institutions, and outputs further evolves as systems become increasingly embedded into daily life. The increased ease of access to generating content and potential misinformation, and difficulty distinguishing between human and AI-generated content, poses risks to trust in media and content authenticity [2].

4.1.1 Trust in Media and Information

Contributors: *Jessica Newman, Irene Solaiman, Canyu Chen, Arjun Subramonian*

The increasing sophistication of generative AI has expanded the possibilities of misleading content and disinformation campaigns [71] and made it harder for people to trust information [57]. AI-generated misinformation [139] can be perpetuated by reinforcement and volume [267] when widely distributed online. Real-world impacts can include loss of trust in mainstream news [143]. LLM-generated misinformation can be harder to detect accurately than human-generated misinformation with the same semantics, indicating it can have more deceptive styles [70]. GPT-3-level systems can also produce more compelling disinformation [327]. Moreover, automated detection systems show flaws, such as often incorrectly flagging non-native language speakers' writing as machine-generated [207]. Issues highlighted in Section 3.3 can contribute to misleading information among subpopulations.

4.1.1.1 What to Evaluate Surveys can examine trust in AI systems [152, 272] to output factual information; trust in researchers, developers, and organizations developing and deploying AI [232]; mitigation and detection measures [318]; and trust in overall media and how it is distributed [361]. Quantitative surveys of consumers across countries gauging understanding of, satisfaction with, and trust in outputs from generative systems, with the survey conducted multiple times over a given period of time, can give insights to user trust in commercialization [130, 222].

Trust can be measured in the category of information, such as information about democratic and policy institutions [265]. Evaluations and countermeasures of false and misleading information remain challenging. There is no universal agreement about what constitutes misinformation, and much of the research on intervention remains siloed [141].

In text, approaches include examining "adversarial factuality" [335], human preference votes, and leaderboards [75] .

In image and video, approaches include drawing from deepfake detection methods [383] and comparing model architectures' impact on detection [293].

In audio, surveys include a level of trust in the person whose voice is replicated and/or the institution or process that person represents [28].

4.1.1.2 Mitigations and Interventions Interventions include encouraging media users to scrutinize post accuracy before sharing [268], encouraging companies to use crowdsourced fact-checking [132] and digital forensics to detect AI-generated content [110]. However, detection loses accuracy as AI systems become more powerful [298]. Research efforts towards watermarking are ongoing [190] yet can be circumvented by users, as are efforts to mitigate the memorization of undesired concepts learned by diffusion models [123, 150].

Emerging legal and regulatory approaches around the world include the EU AI Act, which requires labeling AI-generated content. Policymakers and developers can also ban use cases where false outputs have the greatest risks. In the U.S., the Department of Commerce is developing guidance for content authentication and watermarking to label AI-generated content, which will be used by Federal agencies. The U. S. Federal Trade Commission is also working to prohibit the use of generative AI for impersonation fraud [112].

4.1.2 Overreliance on Outputs

Contributors: *Jessica Newman, Irene Solaiman*

Overreliance on automation, in general, is a long-studied problem [262], and carries over in novel ways to AI-generated content [266]. People are prone to overestimate and overtrust AI, including AI-generated content, especially when outputs appear authoritative or when people are in time-sensitive situations [58]. This can lead to the spread of biased and inaccurate information [148]. Persistent security vulnerabilities that trick systems into outputting inaccurate information [315] exacerbate the harm of overreliance. LLMs can also exhibit deceptive behavior in certain instances [263].

The study of human-generative AI relationships is nascent but growing, and highlights that the anthropomorphism [2] of these technologies and automation bias [125] may contribute to unfounded trust and reliance [285]. Improving the trustworthiness of AI systems is an important ongoing effort across sectors.[9]

4.1.2.1 What to Evaluate *In text*, conversational interfaces for chatbots can elicit trust and other strong emotions and can be abused to subtly change or manipulate people's behaviors or even encourage self-harm [2]. LLMs can be evaluated for their refusal to generate responses to questions with non-existent concepts or false premises [211]. Here, we consider text and code separately. Although both rely on textual representations, the goals, formatting, and semantics of code and text are very dissimilar and require separate consideration and evaluation.

In code, inaccurate outputs [82] can nullify potential benefits and should be evaluated for their limitations [73] and hazards [186], in addition to categories listed in the Technical Base System section.

4.1.2.2 Mitigations and Interventions Protections include vulnerability disclosure, bug bounties, and AI incident databases. In policy and legislation, components of the EU AI Act may also be helpful, such as requiring labeling of AI-generated content and prohibiting certain kinds of manipulation [107]. The U.S. Federal Trade Commission also protects consumers from false or exaggerated claims about AI products [21].[10]

4.1.3 Personal Privacy and Sense of Self

Contributors: *Jessica Newman, Irene Solaiman, Arjun Subramonian*

[9] See [358, 247].

[10] While the FTC has a mandate to protect consumers against false or exaggerated claims, they may opt not to enforce them [180].

Privacy is linked with autonomy, referring to one's ability to act under self-governance, where lack of privacy can hinder one's ability to act independently. Privacy can protect both powerful institutions and vulnerable peoples, and is interpreted and protected differently by culture and social classes throughout history [237]. Legal definitions and protections vary globally [346, 182] and, when violated, can be distinct from harm [63]. Privacy can refer to content shared, seen, or experienced outside the sphere a person has consented to, or in which they expect it to appear or be inferable [249]. Publicly-available content may have varying privacy considerations [317].

4.1.3.1 What to Evaluate In addition to system-level *Privacy and Data Protection* evaluations, societal impacts [325] and harms [324] from the loss and violation of privacy are difficult to enumerate and evaluate, such as loss of opportunity or reputational damage. Violations can shift in power differentials and to personal expectations of privacy [224] and autonomy. The type of private information violated, such as medical information, can trigger different impacts and responses.

4.1.3.2 Mitigations and Interventions Mitigations first should determine who is responsible for an individual's privacy while recognizing limitations of technical or data literacy. Robust protection requires both individual and collective action [10]. Outside of an individualistic framework, certain rights such as refusal [81] and inclusion also require consideration of individual self-determination.

Technical methods to preserve privacy in a generative AI system [54] cannot guarantee full protection. Upholding privacy regulations requires engagement from multiple affected parties [279] and considering the effectiveness of methods such as opt-outs [59] from data collection [213]. Improving common practices and better global regulation for collecting training data can help. Opt-in approaches can provide better protection [352]. Privacy options for users should ease accessibility [376]. Meaningful consent of data and model subjects is key, as per the EU's General Data Protection Regulation [106].

4.2 Inequality, Marginalization, and Violence

Generative AI systems can exacerbate inequality, as argued in Sections 3.1 and 3.3. When deployed or updated, systems' impacts on people and groups can, directly and indirectly, harm and exploit vulnerable and marginalized groups.

4.2.1 Community Erasure

Contributors: *Dylan Baker, Yacine Jernite, Irene Solaiman, Zeerak Talat*

Biases and safety provisions, such as content moderation, can have unequally distributed costs and benefits and can lead to community erasure [146]. Common methods for harmful content removal can lower system performances for marginalized communities [382] and suppress community-specific and reclaimed language [77]. Automated removal can perform poorly or be harmful to marginalized populations [303, 338].

Mitigations often combine four methods: data sourcing [38]; human moderation of content included in training data [87]; automated moderation of content included in training data [147]; and keyword deny-lists [300]. Given that the exclusion of harmful content within datasets stand to create distinct harms to marginalized communities [99, 259], efforts towards mitigation of generating harmful content becomes a question of the politics of classification [50, 200, 100, 350] and its potential harms.

4.2.1.1 What to Evaluate Evaluating *Disparate Performance* once systems have undergone safety provisions can give signal to possible erasure. Accounting for the demographics and composition of human crowdworkers can also provide information [304] about subsequent impacts. Longer-term impacts of erasure depend on the system's deployment context, leading to opportunity loss or reinforced biases and norms.

4.2.1.2 Mitigations and Interventions Better democratic processes for developing and deploying systems and safety provisions such as content moderation should work with marginalized populations. This should include more investment in representative crowdworkers and appropriate compensation and mental health support. Lessons from social media content moderation can apply, such as working with groups who have been erased and documenting patterns of erasure to improve future approaches [312].

4.2.2 Long-term Amplification and Embedding of Marginalization by Exclusion (and Inclusion)

Contributors: *Yacine Jernite, Irene Solaiman, Zeerak Talat*

Inclusion without consent can also harm marginalized groups, including via surveillance and exploitation. For example, while some research strives for greater inclusion of underrepresented and Indigenous languages [167], Indigenous groups have resisted AI approaches [84]. Work conducted on low resource and Indigenous cultural artifacts, such as language and symbolism, should ensure meaningful inclusion of the proprietors and community members [306].

4.2.2.1 Disparate Performance in Critical Infrastructure Generative AI use in critical infrastructure directly impacting human wellbeing can also be classified as high-risk use cases. This includes use in financial services, healthcare such as mental health and medical advice, and democratic processes, such as election or political information. Examples include crisis intervention, as well as research [118] and action [238] to use chatbots for eating disorder prevention. Technical tooling used in human systems and processes that have long-recorded discrimination patterns [374] can instead exacerbate harm [197].

Recent research highlights a disconnect between static fairness objectives and their long-term impacts [133]. For instance, in lending, algorithms can reinforce stigmatization and worsen marginalization over time [212]. Understanding these long-term effects requires considering complex interactions between automated choices, individual responses, and societal dynamics.

4.2.2.2 What to Evaluate Systems should again undergo *Disparate Performance* evaluations once updated for a high-risk task in critical infrastructure and account for overall deployment context. Evaluating marginalization will depend on context, and should account for marginalization when work about and by marginalized populations is less visible or uncredited [377]. Evaluating marginalization impacts on individuals, such as through health [25], is ongoing research.

4.2.2.3 Mitigations and Interventions Information about disparate performance can improved through better evaluation work for underrepresented populations and low-resource languages as well as crediting and including local researchers [45] from these communities.

Engagement with populations should be done in ways that embody local approaches [51]. Policies should be crafted to better respect rights of refusal [320], and which aim to mitigate the common power disparities between model builders and these communities. Nations that address these discriminatory patterns through regulations should coordinate with other nations to promote broad and international protections where possible.

4.2.3 Abusive or Violence Content

Contributors: *Irene Solaiman, Zeerak Talat*

Generative AI systems can generate outputs that are used for abuse, constitute non-consensual content, or are threats of violence and harassment [68]. Non-consensual sexual representations of people, include representations of minors and child sexual abuse material (CSAM) [240]. Abuse and violence can disparately affect groups, such as women and girls[108, 155] These harms are additionally experienced disproportionately on the basis of gender, race, ethnicity, sexual orientation, religion, and other social categories [80].

4.2.3.1 What to Evaluate Sensitive topics and trauma's impacts on people are by nature challenging to evaluate and should be done with care. Consequences of abuse of children and minors can be long-term or lifelong [8]. Impacts and trauma can resurface throughout a person's life in many aspects. Evaluations for generative AI impacts can overlap with similar harms such as image-based sexual abuse [177]. As seen in Section 3.2, consent from affected people should be evaluated with the person themselves.

4.2.3.2 Mitigations and Interventions Research to detect, mitigate, and report abusive and violent content such as CSAM is ongoing [349] and tools specific to modalities such as images can help identify content that is not yet labeled as CSAM [351]. Additionally, datasets should be flagged for containing abusive and violent content, and appropriately not be distributed or used for model training [40, 348]. Relevant regulation should be updated to address generated content that may not accurately portray an existing person or their body or self, but lead to real harms. Institutions such as the Canadian Centre for Child Protection (C3P) and the Internet Watch Foundation (IWF) are dedicated to evaluating for CSAM with relevant legal context. Furthermore, training evaluation models on CSAM or having CSAM reference content for evaluation is also often illegal.

4.3 Concentration of Authority

The concentration of power and authority in decisions about and access to generative AI occurs in numerous interrelated and simultaneous, but not always straightforward ways. Concentrating authoritative power can exacerbate inequality and lead to exploitation.

Few countries, companies, and organizations currently have the ability to develop advanced AI systems [340], and the costs of training generative systems are high, which has caused a shift towards market concentration [193]. Greater transparency across many different indicators is one way that researchers hope to counteract some of the negative effects of the concentration of authority for public accountability [47].

4.3.1 Militarization, Surveillance, and Weaponization

Contributors: *Irene Solaiman, Jessica Newman*

Concentrating power can occur at various levels, from small groups to national bodies. National level power includes surveillance, and interest in the militarization of generative AI systems is growing [151]. Use includes generating synthetic data for training AI systems [359] and military planning [113]. Military use is not inherently weaponization and risk depends on the use case and government interest. AI deployed for national security interests require differentiating national security interests from undue harm [60].

Generative AI systems are also enabling new kinds of cyberattacks, and amplifying the possibilities of existing cyberattacks. For example, synthetic audio has been used for more compelling fraud and extortion [181]. LLMs are also facilitating disinformation campaigns, influence operations, and phishing attacks [138]. Research shows Russian military intelligence actors, North Korean threat actors, Iranian threat actors, and Chinese state-affiliated threat actors are using LLMs to enhance their efforts for reconnaissance, social engineering, and cyber operations [229].

4.3.1.1 What to Evaluate If deployed covertly, under NDA, or without transparency, generative AI systems used for surveillance or weaponization can be difficult to track or evaluate. Evaluations can broadly analyze the quantity of where such systems have been deployed, such as the number of devices sold, or number of system deployments, as a brute force measure. AI developers can also study and monitor how threat actors use generative AI systems to better evade detection or expand malicious capabilities. There are a small number of evaluations that test the ability of LLMs to help carry out cyber attacks [271] or help develop chemical or biological weapons [205].

4.3.1.2 Mitigations and Interventions For release or procurement of technical systems, developers can restrict surveillance and weaponization as use cases [191]. Similarly, academic venues can adopt codes of ethics for weaponization or violating human rights using AI [246]. Government development of generative AI systems for surveillance and weaponization requires additional protocols. Governments and militaries can make commitments toward ethical and responsible uses of AI [92] and joint commitments from multiple countries [360, 230] can create accountability among military powers. Regulatory approaches can draw boundaries for harmful uses by militaries, but will grapple with tensions on what constitutes national security [379], operating within the framework of International Humanitarian Law [44] with respect to autonomous weapons, and moving forward on international agreements on autonomous weapons [161].

For organizations to protect themselves against cyberattacks that make use of generative AI technologies, standard cybersecurity practices such as multifactor authentication and Zero Trust defenses can be helpful. AI developers should also continuously monitor their AI systems to understand and block malicious uses and attacks.

4.3.2 Imposing Norms and Values

Contributors: *Dylan Baker, Yacine Jernite, Marie-Therese Png, Irene Solaiman, Zeerak Talat*

Global deployment of a model can consolidate power within a single, originating culture, to determine and propagate acceptability across cultures [39, 233, 353]. Highest performing characteristics of generative systems such as language, dominant cultural values, and embedded norms can overrepresent regions outside of where a system is deployed. For example, a language model that is highest performing in the English language can be deployed in a region with a different dominant language and incentivize engaging in English, further excluding those who do not have an English language background. Establishing or reinforcing goodness with certain languages, dialects, accents, imagery, social norms, and other representations of peoples and cultures can contribute to these norms and values imposition.

4.3.2.1 What to Evaluate In addition to evaluations and limitations outlined in Section 3.2, complex, qualitative, and evolving cultural concepts such as beauty and success are viewed differently in context of an application and cultural region. The impacts of norm and value impositions are already manifesting [165] and require critical foresight as they evolve. Imposition contributes to homogenization, including the suppression of marginalized identities [94], languages, cultural practices, and epistemologies [221].

4.3.2.2 Mitigations and Interventions Mitigations should be cognizant of preserving irreducible differences among cultures [105] and practicing value-sensitive design [120], including by focusing on system components such as data extraction and use [85]. Methods for cultural value alignment [322] can improve and require improving methods and infrastructure for working with underrepresented groups. Novel alignment techniques by modality can determine preferable principles [372] and values [27] for generative AI systems. Prominent AI regulations should account for "copycat" legislation in other countries.

4.4 Labor and Creativity

AI systems deployed as tools and assistants for human labor, thought, and creativity, should be evaluated for the ongoing effects generative AI systems have on skills, jobs, and the labor market.

4.4.1 Intellectual Property and Ownership

Contributors: *Irene Solaiman, Yacine Jernite*

Rights to the training data [53, 220] and replicated or plagiarized work [52] are ongoing legal and policy discussions [366], often by specific modality. Impacts of generated content to people and society, such as reputational damage and economic loss [172], will necessarily coexist with impacts and development of intellectual property law.

4.4.1.1 What to Evaluate Determining whether original content has been used in training data depends on developer transparency or research on training data extraction [66]. Given the large sizes of training datasets, possible methods of evaluating original content inclusion could be through search and matching tools [104, 326]. In addition to unclear legal implications, the ambiguity of impacts on content ownership [102] makes evaluation difficult.

In image, surveys can examine AI authorship of generated imagery and measure user sentiments towards AI imagery generally [76].

4.4.1.2 Mitigations and Interventions Similar to Personal Privacy and Sense of Self (see Section 4.1.3), opt-in and opt-out mechanisms can protect intellectual property but depend on adherence.

Regulation and stricter rules from a developer organization about training material will differ by modality. Ongoing lawsuits set legal precedent [72]. Tools are being developed to protect certain modalities from being used nonconsensually as training data [345].

4.4.2 Economy and Labor Market

Contributors: *Alberto Lusoli*

Key considerations about the impact of automation and AI on employment center on whether these technologies will generate new jobs or, in contrast, will lead to a large-scale worker displacement in the next future [316]. Some suggest productive supplementing of repetitive tasks [13] while others warn of displacement and polarization [3]. Automation unevenly affects workers, since efficiency can be measured and prioritized disparately [19]. Long-term, research shows how technological advancements have historically increased earning inequality between education, sex, race, and age groups [3]. Market incentives and value attributed to varying skills and may not accurately reflect societal needs, and are often based on gendered and racialized preconceptions of the value of labor [19].

4.4.2.1 What to Evaluate *In text*, approaches to examine short-term effects of LLMs on productivity include measuring a selected group of people using an LLM to supplement a given task, where productivity is measured as earnings per minute. This factors in time for a task and the quality of the output [251].

Across modalities, substitution of labor for capital might cut costs in the short term [79]. For specific tasks, evaluating the quality of generated output compared to human output can signal the likelihood of a generative AI system replacing human labor [309]. The long-term impact on the global economy is unclear and depends on industry decisions. Potential evaluation variables include unemployment rates, salaries for a given skill or task, economic class divisions, and overall cost of services. Positive externalities[11] could stimulate competition, drive prices down, and have a net-positive effect on employment [201]. A task-polarization model [22] shows how AI can potentially widen the gap between high and low-wage occupations at the expense of the middle tier. Further evaluations can investigate types of jobs created and sunsetted.

See Section 3.7 for evaluating human labor in the research, development, and deployment process.

4.4.2.2 Mitigations and Interventions Workers affected by AI can be supported via re-skilling and upskilling opportunities. This will also help reduce the barriers to entry to new jobs. Managing the transition can be challenging and require policy intervention. Prominent movements and worker disapproval, such as in the film and entertainment industry [30], can set precedence. In addition to labor protection laws, more inclusive design processes can open technological decisions to democratic participation and steer innovation in socially desirable directions.

Proposed policy interventions include an "automation tax" [256] to compensate for negative externalities (i.e., unemployment). In practice, an automation tax could be determined by the extent to which layoffs can be attributed to automation, or by measuring the intensity of automation, and adjusting the amount each firm contributes in unemployment insurance payments accordingly. Taxes can generate revenue to support re-skilling programs and slow the introduction of employment-substituting technologies, providing governments with more time to prepare for the potential effects of structural under- and unemployment. Limitations include difficulty in clearly identifying labor-saving from labor-enhancing technologies, the risk of imposing double taxation on capital investments, and possibly slowing technological innovation, GDP and wage growth [20].

4.5 Ecosystem and Environment

Impacts at a high-level, from the AI ecosystem to the Earth itself, are necessarily broad but can be broken down into components for evaluation.

[11] We broadly understand externalities as the unanticipated effects of economic activities on the social environment.

4.5.1 Widening Resource Gaps

Contributors: *Irene Solaiman*

As described in Section 3.6, the high financial and resource costs necessarily excludes groups who do not have the resources to train, evaluate, or host models. The infrastructure needed to contribute to generative AI research and development leads to widening gaps which are notable among sectors, such as between industry and academia [198, 329], or among global powers and countries [11].

4.5.1.1 Access and Benefit Distribution Ability to contribute to and benefit from a system depends on ability to engage with a system, which in turn depends on the openness of the system, the system application, and system interfaces. Level of openness and access grapples with tensions of misuse and risk. Increasing trends toward system closedness [321] is shifting access distribution.

4.5.1.2 Geographic and Regional Activity Concentration In the field of AI as a whole historically, top AI research institutions from 1990-2014 have concentrated in the U.S. [250] More recent data highlights the U.S., EU, and China as primary hubs [294]. Even within the U.S., AI activity concentrates in urban, coastal areas [239].

4.5.1.3 What to Evaluate Evaluation should first determine AI-specific resources, then track trends by sector and region. To determine and evaluate level of access, first components of access should be established. This includes technical details, upstream decisions, auditing access, and opt-out or opt-in reliability. Specific resources such as compute power [6] are popularly tracked by annual reports on the field of AI [32, 329].

4.5.1.4 Mitigations and Interventions Policymakers can minimize resource gaps by making high-cost resources, such as compute power, accessible via applications and grants to researchers and low-resource organizations. Intercultural dialogues [64] that meaningfully address power imbalances and lowering the barrier for underrepresented peoples to contribute can improve harms from resource gaps. This can include accessible interfaces to interact with and conduct research on generative AI systems and low- to no-code tooling.

4.5.2 Environmental Impacts

Contributors: *Sasha Luccioni, Irene Solaiman, Marie-Therese Png, Michelle Lin*

In addition to the *Environmental Impacts* and carbon emissions from a system itself, evaluating impact on the Earth can follow popular frameworks and analyses.

4.5.2.1 What to Evaluate Environmental, social, and governance (ESG) frameworks and the Scope 1, 2, and 3 system can give structure to how developers track carbon emissions [288]. Scope 3 emissions, the indirect emissions often outside a developer's control, should account for a generative AI system's lifecycle, including in deployment [216]. Long-term effects of AI environmental impacts on the world and people can range from inequity to quality of life [287]. Negative environmental impacts of AI are unevenly distributed globally, with mining, data center water consumption, and carbon emissions disproportionately affecting the Global South [115, 209, 236]. Research to measure overall impacts of climate change is ongoing [358].

4.5.2.2 Mitigations and Interventions Systemic change can improve energy and carbon efficiency in ML systems, from energy efficient default settings for platforms and tools, to an awareness of balancing gains with cost; for example, weighing energy costs, both social and monetary, with the performance gains of a new model before deploying it. Regulatory proposals move toward mixed strategies for sustainable AI, including sustainability by design and consumption caps. Best practices for developers and researchers include choosing efficient testing environments, promoting reproducibility, and standardized reporting. An energy efficiency leaderboard can incentivize sustainable research [149].

Antitrust overlap with environmental costs associated with the AI compute stack underscores a complex interplay between market concentration, technological advancement, and environmental

sustainability. Vertical integration observed in the AI compute market has significant implications for the environmental footprint of AI systems [31] when proprietary hardware and software configurations may not be optimized for energy efficiency. By mandating interoperability and separating compute hardware from software, regulatory measures could encourage the adoption of more energy-efficient technologies across the AI lifecycle. Applying antitrust principles to AI compute markets could incentivize greener technologies to attract environmentally conscious consumers and comply with sustainability standards. Antitrust considerations include intervention through merger enforcement and tackling anti-competitive conduct.

Standards and transparency for carbon emissions reporting and accounting for efficiency can help better understand evolution and compare the emissions of different approaches and models. While certain conferences such as NeurIPS are starting to include compute information in submissions as checklists, reporting and figures can vary widely depending on what factors are included. Accuracy may trade off with efficiency. Incentivizing and including these metrics when comparing two or more models (e.g., in benchmarks and leaderboards) can help users make trade-offs that consider both aspects and choose the model that best corresponds to their use case and criteria.

Legislative approaches emphasize the urgent need for comprehensive studies on the AI environmental impacts of artificial intelligence [223].

Broader Impacts and Future Work

Understanding an AI system from conception to training to deployment requires insight into training data, the model itself, and the use case/application into which the system is deployed. It also requires understanding people, society, and how societal processes, institutions, and power are changed and shifted by an AI system.

5 Lack of context for base model

Context is critical to robust evaluation; the way in which we define and evaluate harm in any given application requires an understanding of the target industry, task, end-user, and model architecture. Communication across model developers, model deployers, and end-users is key to developing a comprehensive evaluation and risk mitigation strategy. Actors across the ecosystem should collaborate to craft robust evaluations and invest in the safeguards needed to prevent harm.

6 Context of the Evaluation

Systems can be deployed in contexts where there is not sufficient attention towards evaluating and moderating performance. This means disparate performance is not caught, as seen with social media platform moderation outside of the most commonly-written languages and wealthiest countries [297]. Moreover, as cultural values change between cultural contexts, both within and outside of any given language, the particular cultural values that are being evaluated should be made explicit. A byproduct of such specificity is that it becomes clear where evaluations should be extended while providing a framework for such extensions.

7 Choosing Evaluations

Further work is needed to compare, select, and document evaluations. The evaluations selected to determine a model's performance will impact the values it propagates out during deployment. There is no universal evaluation by which to determine a model's performance, and any evaluation metrics should be used with deployment context in mind [282, 305].Evaluations themselves require further scrutiny and evaluation, to be able to determine the most appropriate evaluations to run. Appropriate evaluations per category should be standardized through policy bodies and coordinated internationally. Furthermore, notable work at top AI ethics publication venues has not adequately centered on the least powerful in society [41].

Conclusion

Just as generative AI systems undergo performance evaluations, they must also be evaluated for social impacts. The seven categories in our framework for technical base systems move toward a standard evaluation framework for listed modalities of a base system. Our analyses of popular evaluation methods per category can help to improve research in producing novel evaluations. Evaluations under the "People and Society" category overlaps with existing risk and harms taxonomies for generative AI systems. The latter evaluation category has limited case studies and must consider challenges and ethics of determining human responses. Since social impact evaluations can only give limited information about each impact type, we recommend that all categories are given equal importance, and that all relevant stakeholders are meaningfully consulted throughout the development, evaluation, and deployment processes.

Acknowledgements

We would like to thank Adina Williams, Levent Sagun, Kevin Klyman, and Apostol Vassilev for their input to drafts of this chapter.

Authorship by Section

Bias, Stereotypes, and Representational Harms: Yacine Jernite, Lama Ahmad, Sara Hooker, Irene Solaiman, Zeerak Talat, Margaret Mitchell, Usman Gohar, Jennifer A Mickel, Dylan Baker, Alina Leidinger, Felix Friedrich, Anaelia Ovalle, Avijit Ghosh, Hal Daumé III

Cultural Values and Sensitive Content: Irene Solaiman, Zeerak Talat, Alina Leidinger, Isabella Duan, Margaret Mitchell, Felix Friedrich, Jennifer Mickel

Disparate Performance: Yacine Jernite, Irene Solaiman, Usman Gohar, Jennifer Mickel, Margaret Mitchell, Arjun Subramonian, Anaelia Ovalle, Avijit Ghosh, Hal Daumé III

Environmental Costs and Carbon Emissions: Sasha Luccioni, Marie-Therese Png, Irene Solaiman, Usman Gohar, Michelle Lin

Privacy and Data Protection: Ellie Evans, Yacine Jernite, Irene Solaiman, Canyu Chen, Jessica Newman, Shubham Singh, Isabella Duan, Lukas Struppek, Arjun Subramonian

Financial Costs: Irene Solaiman, Anaelia Ovalle

Data and Content Moderation Labor: Dylan Baker, Yacine Jernite, Alberto Lusoli, Irene Solaiman, Jennifer Mickel, Arjun Subramonian

Trust in Media and Information: Jessica Newman, Irene Solaiman, Canyu Chen, Arjun Subramonian

Overreliance on Outputs: Jessica Newman, Irene Solaiman

Personal Privacy and Sense of Self: Jessica Newman, Irene Solaiman, Arjun Subramonian

Community Erasure: Dylan Baker, Yacine Jernite, Irene Solaiman, Zeerak Talat

Long-term Amplification and Embedding of Marginalization by Exclusion (and Inclusion): Yacine Jernite, Irene Solaiman, Zeerak Talat

Abusive or Violence Content: Irene Solaiman, Zeerak Talat

Militarization, Surveillance, and Weaponization: Irene Solaiman, Jessica Newman

Imposing Norms and Values: Dylan Baker, Yacine Jernite, Marie-Therese Png, Irene Solaiman, Zeerak Talat

Intellectual Property and Ownership: Irene Solaiman, Yacine Jernite

Economy and Labor Market: Alberto Lusoli, Irene Solaiman

Widening Resource Gaps: Irene Solaiman

Environmental Impacts: Sasha Luccioni, Irene Solaiman, Marie-Therese Png, Michelle Lin

References

[1] L. Abbott and C. Grady. A Systematic Review of the Empirical Literature Evaluating IRBs: What We Know and What We Still Need to Learn. *Journal of Empirical Research on Human Research Ethics*, 6(1):3–19, Mar. 2011. ISSN 1556-2646, 1556-2654. doi: 10.1525/jer.2011.6.1.3. URL http://journals.sagepub.com/doi/10.1525/jer.2011.6.1.3.

[2] G. Abercrombie, A. Curry, T. Dinkar, V. Rieser, and Z. Talat. Mirages. On Anthropomorphism in Dialogue Systems. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 4776–4790, Singapore, 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.290. URL https://aclanthology.org/2023.emnlp-main.290.

[3] D. Acemoglu and P. Restrepo. Tasks, Automation, and the Rise in U.S. Wage Inequality. *Econometrica*, 90(5):1973–2016, 2022. ISSN 0012-9682. doi: 10.3982/ECTA19815. URL https://www.econometricsociety.org/doi/10.3982/ECTA19815.

[4] O. Ahia, J. Kreutzer, and S. Hooker. The Low-Resource Double Bind: An Empirical Study of Pruning for Low-Resource Machine Translation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3316–3333, Punta Cana, Dominican Republic, 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.findings-emnlp.282. URL https://aclanthology.org/2021.findings-emnlp.282.

[5] O. Ahia, S. Kumar, H. Gonen, J. Kasai, D. Mortensen, N. Smith, and Y. Tsvetkov. Do All Languages Cost the Same? Tokenization in the Era of Commercial Language Models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9904–9923, Singapore, 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.614. URL https://aclanthology.org/2023.emnlp-main.614.

[6] N. Ahmed and M. Wahed. The De-democratization of AI: Deep Learning and the Compute Divide in Artificial Intelligence Research, Oct. 2020. URL http://arxiv.org/abs/2010.15581.

[7] O. Aka, K. Burke, A. Bauerle, C. Greer, and M. Mitchell. Measuring Model Biases in the Absence of Ground Truth. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, pages 327–335, Virtual Event USA, July 2021. ACM. ISBN 978-1-4503-8473-5. doi: 10.1145/3461702.3462557. URL https://dl.acm.org/doi/10.1145/3461702.3462557.

[8] A. Al Odhayani, W. J. Watson, and L. Watson. Behavioural consequences of child abuse. *Canadian Family Physician*, 59(8):831–836, Aug. 2013. ISSN 0008-350X. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3743691/.

[9] All Tech Is Human. AI and Human Rights: Building a Tech Future Aligned With the Public Interest. Technical report, All Tech Is Human, New York, NY, USA, 2022. URL https://alltechishuman.org/ai-human-rights-report.

[10] A. L. Allen. Protecting One's Own Privacy in a Big Data Economy. *Harvard Law Review*, 130 (2), 2016. URL https://harvardlawreview.org/forum/vol-130/protecting-ones-own-privacy-in-a-big-data-economy/.

[11] C. Alonso, S. Kothari, and S. Rehman. How Artificial Intelligence Could Widen the Gap Between Rich and Poor Nations, Dec. 2020. URL https://www.imf.org/en/Blogs/Articles/2020/12/02/blog-how-artificial-intelligence-could-widen-the-gap-between-rich-and-poor-nations.

[12] Amazon Web Services. Amazon S3 Simple Storage Service Pricing. URL https://aws.amazon.com/s3/pricing/.

[13] N. Anantrasirichai and D. Bull. Artificial intelligence in the creative industries: A review. *Artificial Intelligence Review*, 55(1):589–656, Jan. 2022. ISSN 1573-7462. doi: 10.1007/s10462-021-10039-7. URL https://doi.org/10.1007/s10462-021-10039-7.

[14] C. Andrade. The Limitations of Online Surveys. *Indian Journal of Psychological Medicine*, 42(6):575–576, Nov. 2020. ISSN 0253-7176, 0975-1564. doi: 10.1177/0253717620957496. URL http://journals.sagepub.com/doi/10.1177/0253717620957496.

[15] M. Andrus, E. Spitzer, J. Brown, and A. Xiang. What We Can't Measure, We Can't Understand: Challenges to Demographic Data Procurement in the Pursuit of Fairness. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 249–260, Virtual Event Canada, Mar. 2021. ACM. ISBN 978-1-4503-8309-7. doi: 10.1145/3442188.3445888. URL https://dl.acm.org/doi/10.1145/3442188.3445888.

[16] L. F. W. Anthony, B. Kanding, and R. Selvan. Carbontracker: Tracking and Predicting the Carbon Footprint of Training Deep Learning Models, July 2020. URL http://arxiv.org/abs/2007.03051.

[17] Anthropic. Introducing the next generation of Claude, 2024. URL https://www.anthropic.com/news/claude-3-family.

[18] C. Ashurst, E. Hine, P. Sedille, and A. Carlier. AI Ethics Statements: Analysis and Lessons Learnt from NeurIPS Broader Impact Statements. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 2047–2056, Seoul Republic of Korea, June 2022. ACM. ISBN 978-1-4503-9352-2. doi: 10.1145/3531146.3533780. URL https://dl.acm.org/doi/10.1145/3531146.3533780.

[19] N. Atanasoski and K. Vora. *Surrogate Humanity: Race, Robots, and the Politics of Technological Futures*. Perverse Modernities. Duke University Press, Durham, NC, 2019. ISBN 978-1-4780-0386-1.

[20] R. D. Atkinson. The Case Against Taxing Robots, Apr. 2019. URL https://papers.ssrn.com/abstract=3382824.

[21] M. Atleson. Keep your AI claims in check, Feb. 2023. URL https://www.ftc.gov/business-guidance/blog/2023/02/keep-your-ai-claims-check.

[22] D. Autor. The Labor Market Impacts of Technological Change: From Unbridled Enthusiasm to Qualified Optimism to Vast Uncertainty, May 2022. URL https://www.nber.org/papers/w30074.

[23] S. K. B, A. Chandrabose, and B. R. Chakravarthi. An Overview of Fairness in Data – Illuminating the Bias in Data Pipeline. In B. R. Chakravarthi, J. P. McCrae, M. Zarrouk, K. Bali, and P. Buitelaar, editors, *Proceedings of the First Workshop on Language Technology for Equality, Diversity and Inclusion*, pages 34–45, Kyiv, Apr. 2021. Association for Computational Linguistics. URL https://aclanthology.org/2021.ltedi-1.5.

[24] B-Tech and United Nations Human Rights Office of the High Commissioner. Taxonomy of Human Rights Risks Connected to Generative AI. Technical report, United Nations Human Rights Office of the High Commissioner, 2023. URL https://www.ohchr.org/sites/default/files/documents/issues/business/b-tech/taxonomy-GenAI-Human-Rights-Harms.pdf.

[25] F. O. Baah, A. M. Teitelman, and B. Riegel. Marginalization: Conceptualizing patient vulnerabilities in the framework of social determinants of health – An integrative review. *Nursing inquiry*, 26(1):e12268, Jan. 2019. ISSN 1320-7881. doi: 10.1111/nin.12268. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6342665/.

[26] E. Bagdasaryan, O. Poursaeed, and V. Shmatikov. Differential privacy has disparate impact on model accuracy. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*. Curran Associates Inc., Red Hook, NY, USA, 2019.

[27] Y. Bai, S. Kadavath, S. Kundu, A. Askell, J. Kernion, A. Jones, A. Chen, A. Goldie, A. Mirhoseini, C. McKinnon, C. Chen, C. Olsson, C. Olah, D. Hernandez, D. Drain, D. Ganguli, D. Li, E. Tran-Johnson, E. Perez, J. Kerr, J. Mueller, J. Ladish, J. Landau, K. Ndousse, K. Lukosuite, L. Lovitt, M. Sellitto, N. Elhage, N. Schiefer, N. Mercado, N. DasSarma, R. Lasenby, R. Larson, S. Ringer, S. Johnston, S. Kravec, S. E. Showk, S. Fort, T. Lanham, T. Telleen-Lawton,

T. Conerly, T. Henighan, T. Hume, S. R. Bowman, Z. Hatfield-Dodds, B. Mann, D. Amodei, N. Joseph, S. McCandlish, T. Brown, and J. Kaplan. Constitutional AI: Harmlessness from AI Feedback, Dec. 2022. URL http://arxiv.org/abs/2212.08073.

[28] J. Barnett. The Ethical Implications of Generative Audio Models: A Systematic Literature Review. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, pages 146–161, Montr\'{e}al QC Canada, Aug. 2023. ACM. ISBN 9798400702310. doi: 10.1145/3600211.3604686. URL https://dl.acm.org/doi/10.1145/3600211.3604686.

[29] A. M. Barrett, D. Hendrycks, J. Newman, and B. Nonnecke. Actionable Guidance for High-Consequence AI Risk Management: Towards Standards Addressing AI Catastrophic Risks, Feb. 2023. URL http://arxiv.org/abs/2206.08966.

[30] W. Bedingfield. Hollywood Writers Reached an AI Deal That Will Rewrite History, 2023. URL https://www.wired.com/story/us-writers-strike-ai-provisions-precedents/.

[31] H. Belfield and S.-S. Hua. Compute and Antitrust: Regulatory implications of the AI hardware supply chain, from chip design to cloud APIs. *Verfassungsblog*, Aug. 2022. URL https://verfassungsblog.de/compute-and-antitrust/.

[32] N. Benaich, A. Chalmers, O. Sebbouh, and C. Gurau. State of AI Report 2023. Technical report, State of AI, 2023. URL https://www.stateof.ai/.

[33] E. M. Bender, T. Gebru, A. McMillan-Major, and S. Shmitchell. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? 🦜. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 610–623, Virtual Event Canada, Mar. 2021. ACM. ISBN 978-1-4503-8309-7. doi: 10.1145/3442188.3445922. URL https://dl.acm.org/doi/10.1145/3442188.3445922.

[34] C. L. Bennett, C. Gleason, M. K. Scheuerman, J. P. Bigham, A. Guo, and A. To. "It's Complicated": Negotiating Accessibility and (Mis)Representation in Image Descriptions of Race, Gender, and Disability. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI '21, pages 1–19, New York, NY, USA, May 2021. Association for Computing Machinery. ISBN 978-1-4503-8096-6. doi: 10.1145/3411764.3445498. URL https://doi.org/10.1145/3411764.3445498.

[35] J. Berg, M. Furrer, E. Harmon, U. Rani, and M. S. Silberman. Digital labour platforms and the future of work: Towards decent work in the online world. Report, International Labour Organization, Sept. 2018. URL https://apo.org.au/node/244461.

[36] A. Berthelot, E. Caron, M. Jay, and L. Lefèvre. Estimating the environmental impact of Generative-AI services using an LCA-based methodology. In *CIRP LCE 2024 - 31st Conference on Life Cycle Engineering*, pages 1–10, Turin, Italy, June 2024. URL https://inria.hal.science/hal-04346102.

[37] S. Bhatt, S. Dev, P. Talukdar, S. Dave, and V. Prabhakaran. Re-contextualizing fairness in NLP: The case of India. In Y. He, H. Ji, S. Li, Y. Liu, and C.-H. Chang, editors, *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 727–740, Online only, Nov. 2022. Association for Computational Linguistics. URL https://aclanthology.org/2022.aacl-main.55.

[38] BigScience Workshop, T. L. Scao, A. Fan, C. Akiki, E. Pavlick, S. Ilić, D. Hesslow, R. Castagné, A. S. Luccioni, F. Yvon, M. Gallé, J. Tow, A. M. Rush, S. Biderman, A. Webson, P. S. Ammanamanchi, T. Wang, B. Sagot, N. Muennighoff, A. V. del Moral, O. Ruwase, R. Bawden, S. Bekman, A. McMillan-Major, I. Beltagy, H. Nguyen, L. Saulnier, S. Tan, P. O. Suarez, V. Sanh, H. Laurençon, Y. Jernite, J. Launay, M. Mitchell, C. Raffel, A. Gokaslan, A. Simhi, A. Soroa, A. F. Aji, A. Alfassy, A. Rogers, A. K. Nitzav, C. Xu, C. Mou, C. Emezue, C. Klamm, C. Leong, D. van Strien, D. I. Adelani, D. Radev, E. G. Ponferrada, E. Levkovizh, E. Kim, E. B. Natan, F. De Toni, G. Dupont, G. Kruszewski, G. Pistilli, H. Elsahar, H. Benyamina, H. Tran, I. Yu, I. Abdulmumin, I. Johnson, I. Gonzalez-Dios, J. de la Rosa, J. Chim,

J. Dodge, J. Zhu, J. Chang, J. Frohberg, J. Tobing, J. Bhattacharjee, K. Almubarak, K. Chen, K. Lo, L. Von Werra, L. Weber, L. Phan, L. B. allal, L. Tanguy, M. Dey, M. R. Muñoz, M. Masoud, M. Grandury, M. Šaško, M. Huang, M. Coavoux, M. Singh, M. T.-J. Jiang, M. C. Vu, M. A. Jauhar, M. Ghaleb, N. Subramani, N. Kassner, N. Khamis, O. Nguyen, O. Espejel, O. de Gibert, P. Villegas, P. Henderson, P. Colombo, P. Amuok, Q. Lhoest, R. Harliman, R. Bommasani, R. L. López, R. Ribeiro, S. Osei, S. Pyysalo, S. Nagel, S. Bose, S. H. Muhammad, S. Sharma, S. Longpre, S. Nikpoor, S. Silberberg, S. Pai, S. Zink, T. T. Torrent, T. Schick, T. Thrush, V. Danchev, V. Nikoulina, V. Laippala, V. Lepercq, V. Prabhu, Z. Alyafeai, Z. Talat, A. Raja, B. Heinzerling, C. Si, D. E. Taşar, E. Salesky, S. J. Mielke, W. Y. Lee, A. Sharma, A. Santilli, A. Chaffin, A. Stiegler, D. Datta, E. Szczechla, G. Chhablani, H. Wang, H. Pandey, H. Strobelt, J. A. Fries, J. Rozen, L. Gao, L. Sutawika, M. S. Bari, M. S. Al-shaibani, M. Manica, N. Nayak, R. Teehan, S. Albanie, S. Shen, S. Ben-David, S. H. Bach, T. Kim, T. Bers, T. Fevry, T. Neeraj, U. Thakker, V. Raunak, X. Tang, Z.-X. Yong, Z. Sun, S. Brody, Y. Uri, H. Tojarieh, A. Roberts, H. W. Chung, J. Tae, J. Phang, O. Press, C. Li, D. Narayanan, H. Bourfoune, J. Casper, J. Rasley, M. Ryabinin, M. Mishra, M. Zhang, M. Shoeybi, M. Peyrounette, N. Patry, N. Tazi, O. Sanseviero, P. von Platen, P. Cornette, P. F. Lavallée, R. Lacroix, S. Rajbhandari, S. Gandhi, S. Smith, S. Requena, S. Patil, T. Dettmers, A. Baruwa, A. Singh, A. Cheveleva, A.-L. Ligozat, A. Subramonian, A. Névéol, C. Lovering, D. Garrette, D. Tunuguntla, E. Reiter, E. Taktasheva, E. Voloshina, E. Bogdanov, G. I. Winata, H. Schoelkopf, J.-C. Kalo, J. Novikova, J. Z. Forde, J. Clive, J. Kasai, K. Kawamura, L. Hazan, M. Carpuat, M. Clinciu, N. Kim, N. Cheng, O. Serikov, O. Antverg, O. van der Wal, R. Zhang, R. Zhang, S. Gehrmann, S. Mirkin, S. Pais, T. Shavrina, T. Scialom, T. Yun, T. Limisiewicz, V. Rieser, V. Protasov, V. Mikhailov, Y. Pruksachatkun, Y. Belinkov, Z. Bamberger, Z. Kasner, A. Rueda, A. Pestana, A. Feizpour, A. Khan, A. Faranak, A. Santos, A. Hevia, A. Unldreaj, A. Aghagol, A. Abdollahi, A. Tammour, A. HajiHosseini, B. Behroozi, B. Ajibade, B. Saxena, C. M. Ferrandis, D. McDuff, D. Contractor, D. Lansky, D. David, D. Kiela, D. A. Nguyen, E. Tan, E. Baylor, E. Ozoani, F. Mirza, F. Ononiwu, H. Rezanejad, H. Jones, I. Bhattacharya, I. Solaiman, I. Sedenko, I. Nejadgholi, J. Passmore, J. Seltzer, J. B. Sanz, L. Dutra, M. Samagaio, M. Elbadri, M. Mieskes, M. Gerchick, M. Akinlolu, M. McKenna, M. Qiu, M. Ghauri, M. Burynok, N. Abrar, N. Rajani, N. Elkott, N. Fahmy, O. Samuel, R. An, R. Kromann, R. Hao, S. Alizadeh, S. Shubber, S. Wang, S. Roy, S. Viguier, T. Le, T. Oyebade, T. Le, Y. Yang, Z. Nguyen, A. R. Kashyap, A. Palasciano, A. Callahan, A. Shukla, A. Miranda-Escalada, A. Singh, B. Beilharz, B. Wang, C. Brito, C. Zhou, C. Jain, C. Xu, C. Fourrier, D. L. Periñán, D. Molano, D. Yu, E. Manjavacas, F. Barth, F. Fuhrimann, G. Altay, G. Bayrak, G. Burns, H. U. Vrabec, I. Bello, I. Dash, J. Kang, J. Giorgi, J. Golde, J. D. Posada, K. R. Sivaraman, L. Bulchandani, L. Liu, L. Shinzato, M. H. de Bykhovetz, M. Takeuchi, M. Pàmies, M. A. Castillo, M. Nezhurina, M. Sänger, M. Samwald, M. Cullan, M. Weinberg, M. De Wolf, M. Mihaljcic, M. Liu, M. Freidank, M. Kang, N. Seelam, N. Dahlberg, N. M. Broad, N. Muellner, P. Fung, P. Haller, R. Chandrasekhar, R. Eisenberg, R. Martin, R. Canalli, R. Su, R. Su, S. Cahyawijaya, S. Garda, S. S. Deshmukh, S. Mishra, S. Kiblawi, S. Ott, S. Sang-aroonsiri, S. Kumar, S. Schweter, S. Bharati, T. Laud, T. Gigant, T. Kainuma, W. Kusa, Y. Labrak, Y. S. Bajaj, Y. Venkatraman, Y. Xu, Y. Xu, Y. Xu, Z. Tan, Z. Xie, Z. Ye, M. Bras, Y. Belkada, and T. Wolf. BLOOM: A 176B-Parameter Open-Access Multilingual Language Model, June 2023. URL http://arxiv.org/abs/2211.05100.

[39] A. Birhane and Z. Talat. It's incomprehensible: On machine learning and decoloniality. In S. Lindgren, editor, *Handbook of Critical Studies of Artificial Intelligence*, pages 128–140. Edward Elgar Publishing, Nov. 2023. ISBN 978-1-80392-856-2 978-1-80392-855-5. doi: 10.4337/9781803928562.00016. URL https://www.elgaronline.com/view/book/9781803928562/book-part-9781803928562-16.xml.

[40] A. Birhane, V. U. Prabhu, and E. Kahembwe. Multimodal datasets: Misogyny, pornography, and malignant stereotypes, Oct. 2021. URL http://arxiv.org/abs/2110.01963.

[41] A. Birhane, E. Ruane, T. Laurent, M. S. Brown, J. Flowers, A. Ventresque, and C. L. Dancy. The Forgotten Margins of AI Ethics. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 948–958, Seoul Republic of Korea, June 2022. ACM. ISBN 978-1-4503-9352-2. doi: 10.1145/3531146.3533157. URL https://dl.acm.org/doi/10.1145/3531146.3533157.

[42] A. Birhane, V. Prabhu, S. Han, and V. N. Boddeti. On Hate Scaling Laws For Data-Swamps, June 2023. URL http://arxiv.org/abs/2306.13141.

[43] A. Birhane, V. Prabhu, S. Han, V. N. Boddeti, and A. S. Luccioni. Into the LAIONs Den: Investigating Hate in Multimodal Datasets, Nov. 2023. URL http://arxiv.org/abs/2311.03449.

[44] M. Bo, L. Bruun, and V. Boulanin. Retaining Human Responsibility in the Development and Use of Autonomous Weapon Systems: On Accountability for Violations of International Humanitarian Law Involving AWS. Technical report, Stockholm International Peace Research Institute, Oct. 2022. URL https://www.sipri.org/publications/2022/other-publications/retaining-human-responsibility-development-and-use-autonomous-weapon-systems-accountability.

[45] M. J. Bockarie. We need to end "parachute" research which sidelines the work of African scientists. *Quartz*, Jan. 2019. URL https://qz.com/africa/1536355/african-scientists-are-sidelined-by-parachute-research-teams.

[46] R. Bommasani, D. A. Hudson, E. Adeli, R. Altman, S. Arora, S. von Arx, M. S. Bernstein, J. Bohg, A. Bosselut, E. Brunskill, E. Brynjolfsson, S. Buch, D. Card, R. Castellon, N. S. Chatterji, A. S. Chen, K. A. Creel, J. Davis, D. Demszky, C. Donahue, M. Doumbouya, E. Durmus, S. Ermon, J. Etchemendy, K. Ethayarajh, L. Fei-Fei, C. Finn, T. Gale, L. E. Gillespie, K. Goel, N. D. Goodman, S. Grossman, N. Guha, T. Hashimoto, P. Henderson, J. Hewitt, D. E. Ho, J. Hong, K. Hsu, J. Huang, T. F. Icard, S. Jain, D. Jurafsky, P. Kalluri, S. Karamcheti, G. Keeling, F. Khani, O. Khattab, P. W. Koh, M. S. Krass, R. Krishna, R. Kuditipudi, A. Kumar, F. Ladhak, M. Lee, T. Lee, J. Leskovec, I. Levent, X. L. Li, X. Li, T. Ma, A. Malik, C. D. Manning, S. P. Mirchandani, E. Mitchell, Z. Munyikwa, S. Nair, A. Narayan, D. Narayanan, B. Newman, A. Nie, J. C. Niebles, H. Nilforoshan, J. F. Nyarko, G. Ogut, L. Orr, I. Papadimitriou, J. S. Park, C. Piech, E. Portelance, C. Potts, A. Raghunathan, R. Reich, H. Ren, F. Rong, Y. H. Roohani, C. Ruiz, J. Ryan, C. R'e, D. Sadigh, S. Sagawa, K. Santhanam, A. Shih, K. P. Srinivasan, A. Tamkin, R. Taori, A. W. Thomas, F. Tramèr, R. E. Wang, W. Wang, B. Wu, J. Wu, Y. Wu, S. M. Xie, M. Yasunaga, J. You, M. A. Zaharia, M. Zhang, T. Zhang, X. Zhang, Y. Zhang, L. Zheng, K. Zhou, and P. Liang. On the opportunities and risks of foundation models. *ArXiv*, 2021. URL https://crfm.stanford.edu/assets/report.pdf.

[47] R. Bommasani, K. Klyman, S. Longpre, S. Kapoor, N. Maslej, B. Xiong, D. Zhang, and P. Liang. The Foundation Model Transparency Index, Oct. 2023. URL http://arxiv.org/abs/2310.12941.

[48] R. Bommasani, P. Liang, and T. Lee. Holistic Evaluation of Language Models. *Annals of the New York Academy of Sciences*, 1525(1):140–146, July 2023. ISSN 0077-8923, 1749-6632. doi: 10.1111/nyas.15007. URL https://nyaspubs.onlinelibrary.wiley.com/doi/10.1111/nyas.15007.

[49] L. Bouza, A. Bugeau, and L. Lannelongue. How to estimate carbon footprint when training deep learning models? A guide and review. *Environmental Research Communications*, 5 (11):115014, Nov. 2023. ISSN 2515-7620. doi: 10.1088/2515-7620/acf81b. URL https://iopscience.iop.org/article/10.1088/2515-7620/acf81b.

[50] G. C. Bowker and S. L. Star. *Sorting Things out: Classification and Its Consequences*. Inside Technology. MIT Press, Cambridge, Mass., 1. paperback ed., 8. print edition, 2008. ISBN 978-0-262-52295-3 978-0-262-02461-7.

[51] M. Brereton, P. Roe, R. Schroeter, and A. Lee Hong. Beyond ethnography: Engagement and reciprocity as foundations for design research out here. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '14, pages 1183–1186, New York, NY, USA, Apr. 2014. Association for Computing Machinery. ISBN 978-1-4503-2473-1. doi: 10.1145/2556288.2557374. URL https://doi.org/10.1145/2556288.2557374.

[52] J. Brewster, M. Wang, and C. Palmer. How Copycat Sites Use AI to Plagiarize News Articles. *NewsWeek*, 2023. URL https://www.newsweek.com/how-copycat-sites-use-ai-plagiarize-news-articles-1835212.

[53] B. Brittain. OpenAI hit with new lawsuits from news outlets over AI training. *Reuters*, Feb. 2024. URL https://www.reuters.com/legal/litigation/openai-hit-with-new-lawsuits-news-outlets-over-ai-training-2024-02-28/.

[54] H. Brown, K. Lee, F. Mireshghallah, R. Shokri, and F. Tramèr. What Does it Mean for a Language Model to Preserve Privacy? In *2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 2280–2292, Seoul Republic of Korea, June 2022. ACM. ISBN 978-1-4503-9352-2. doi: 10.1145/3531146.3534642. URL https://dl.acm.org/doi/10.1145/3531146.3534642.

[55] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei. Language models are few-shot learners. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS'20, Red Hook, NY, USA, 2020. Curran Associates Inc. ISBN 978-1-71382-954-6.

[56] M. Brundage, S. Avin, J. Wang, H. Belfield, G. Krueger, G. Hadfield, H. Khlaaf, J. Yang, H. Toner, R. Fong, T. Maharaj, P. W. Koh, S. Hooker, J. Leung, A. Trask, E. Bluemke, J. Lebensold, C. O'Keefe, M. Koren, T. Ryffel, J. B. Rubinovitz, T. Besiroglu, F. Carugati, J. Clark, P. Eckersley, S. de Haas, M. Johnson, B. Laurie, A. Ingerman, I. Krawczuk, A. Askell, R. Cammarota, A. Lohn, D. Krueger, C. Stix, P. Henderson, L. Graham, C. Prunkl, B. Martin, E. Seger, N. Zilberman, S. Ó. hÉigeartaigh, F. Kroeger, G. Sastry, R. Kagan, A. Weller, B. Tse, E. Barnes, A. Dafoe, P. Scharre, A. Herbert-Voss, M. Rasser, S. Sodhani, C. Flynn, T. K. Gilbert, L. Dyer, S. Khan, Y. Bengio, and M. Anderljung. Toward Trustworthy AI Development: Mechanisms for Supporting Verifiable Claims, Apr. 2020. URL http://arxiv.org/abs/2004.07213.

[57] B. Buchanan, A. Lohn, M. Musser, and K. Sedova. Truth, Lies, and Automation. Technical report, Center for Security and Emerging Technology, 2021. URL https://cset.georgetown.edu/publication/truth-lies-and-automation/.

[58] Z. Buçinca, M. B. Malaya, and K. Z. Gajos. To Trust or to Think: Cognitive Forcing Functions Can Reduce Overreliance on AI in AI-assisted Decision-making. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW1):1–21, Apr. 2021. ISSN 2573-0142. doi: 10.1145/3449287. URL http://arxiv.org/abs/2102.09692.

[59] D. Bui, B. Tang, and K. G. Shin. Do Opt-Outs Really Opt Me Out? In *Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security*, CCS '22, pages 425–439, New York, NY, USA, Nov. 2022. Association for Computing Machinery. ISBN 978-1-4503-9450-5. doi: 10.1145/3548606.3560574. URL https://dl.acm.org/doi/10.1145/3548606.3560574.

[60] W. Burke-White. Human Rights and National Security: The Strategic Correlation. *Harvard Human Rights Journal*, Jan. 2004. URL https://scholarship.law.upenn.edu/faculty_scholarship/960.

[61] California Legislative Service. California Consumer Privacy Act of 2018, 2018. URL https://oag.ca.gov/privacy/ccpa.

[62] A. Caliskan, J. J. Bryson, and A. Narayanan. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186, Apr. 2017. ISSN 0036-8075. doi: 10.1126/science.aal4230.

[63] M. Calo. The Boundaries of Privacy Harm. *86 Indiana Law Journal 1131 (2011)*, 86(3), July 2011. ISSN 0019-6665. URL https://www.repository.law.indiana.edu/ilj/vol86/iss3/8.

[64] R. Capurro. The Promising Field of Intercultural Information Ethics. In J. J. Frühbauer, T. Hausmanninger, and R. Capurro, editors, *Localizing the Internet*, pages 9–18. Brill | Fink, Jan. 2018. ISBN 978-3-7705-4200-0 978-3-8467-4200-6. doi: 10.30965/9783846742006_002. URL https://brill.com/view/book/edcoll/9783846742006/BP000002.xml.

[65] N. Carlini, C. Liu, Ú. Erlingsson, J. Kos, and D. Song. The secret sharer: Evaluating and testing unintended memorization in neural networks. In *Proceedings of the 28th USENIX Conference on Security Symposium*, SEC'19, pages 267–284, USA, 2019. USENIX Association. ISBN 978-1-939133-06-9.

[66] N. Carlini, J. Hayes, M. Nasr, M. Jagielski, V. Sehwag, F. Tramèr, B. Balle, D. Ippolito, and E. Wallace. Extracting training data from diffusion models. In *Proceedings of the 32nd USENIX Conference on Security Symposium*, SEC '23, pages 5253–5270, USA, Aug. 2023. USENIX Association. ISBN 978-1-939133-37-3.

[67] N. Carlini, D. Ippolito, M. Jagielski, K. Lee, F. Tramer, and C. Zhang. Quantifying Memorization Across Neural Language Models. In *The Eleventh International Conference on Learning Representations*. ArXiv, Mar. 2023. URL http://arxiv.org/abs/2202.07646.

[68] Center for Technology and Society. Americans' Views on Generative Artificial Intelligence, Hate and Harassment, 2023. URL https://www.adl.org/resources/blog/americans-views-generative-artificial-intelligence-hate-and-harassment.

[69] A. Chan, R. Salganik, A. Markelius, C. Pang, N. Rajkumar, D. Krasheninnikov, L. Langosco, Z. He, Y. Duan, M. Carroll, M. Lin, A. Mayhew, K. Collins, M. Molamohammadi, J. Burden, W. Zhao, S. Rismani, K. Voudouris, U. Bhatt, A. Weller, D. Krueger, and T. Maharaj. Harms from Increasingly Agentic Algorithmic Systems. In *2023 ACM Conference on Fairness, Accountability, and Transparency*, pages 651–666, Chicago IL USA, June 2023. ACM. ISBN 9798400701924. doi: 10.1145/3593013.3594033. URL https://dl.acm.org/doi/10.1145/3593013.3594033.

[70] C. Chen and K. Shu. Can LLM-Generated Misinformation Be Detected? In *The Twelfth International Conference on Learning Representations*, Oct. 2023. URL https://openreview.net/forum?id=ccxD4mtkTU.

[71] C. Chen and K. Shu. Combating Misinformation in the Age of LLMs: Opportunities and Challenges, Nov. 2023. URL http://arxiv.org/abs/2311.05656.

[72] M. Chen. Artists and Illustrators Are Suing Three A.I. Art Generators for Scraping and 'Collaging' Their Work Without Consent, Jan. 2023. URL https://news.artnet.com/art-world/class-action-lawsuit-ai-generators-deviantart-midjourney-stable-diffusion-2246770.

[73] M. Chen, J. Tworek, H. Jun, Q. Yuan, H. P. d. O. Pinto, J. Kaplan, H. Edwards, Y. Burda, N. Joseph, G. Brockman, A. Ray, R. Puri, G. Krueger, M. Petrov, H. Khlaaf, G. Sastry, P. Mishkin, B. Chan, S. Gray, N. Ryder, M. Pavlov, A. Power, L. Kaiser, M. Bavarian, C. Winter, P. Tillet, F. P. Such, D. Cummings, M. Plappert, F. Chantzis, E. Barnes, A. Herbert-Voss, W. H. Guss, A. Nichol, A. Paino, N. Tezak, J. Tang, I. Babuschkin, S. Balaji, S. Jain, W. Saunders, C. Hesse, A. N. Carr, J. Leike, J. Achiam, V. Misra, E. Morikawa, A. Radford, M. Knight, M. Brundage, M. Murati, K. Mayer, P. Welinder, B. McGrew, D. Amodei, S. McCandlish, I. Sutskever, and W. Zaremba. Evaluating Large Language Models Trained on Code, July 2021. URL http://arxiv.org/abs/2107.03374.

[74] M. Cheng, E. Durmus, and D. Jurafsky. Marked Personas: Using Natural Language Prompts to Measure Stereotypes in Language Models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1504–1532, Toronto, Canada, 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.84. URL https://aclanthology.org/2023.acl-long.84.

[75] W.-L. Chiang, L. Zheng, Y. Sheng, A. N. Angelopoulos, T. Li, D. Li, H. Zhang, B. Zhu, M. Jordan, J. E. Gonzalez, and I. Stoica. Chatbot Arena: An Open Platform for Evaluating LLMs by Human Preference, Mar. 2024. URL http://arxiv.org/abs/2403.04132.

[76] S. G. Chiarella, G. Torromino, D. M. Gagliardi, D. Rossi, F. Babiloni, and G. Cartocci. Investigating the negative bias towards artificial intelligence: Effects of prior assignment of AI-authorship on the aesthetic appreciation of abstract paintings. *Computers in Human Behavior*, 137:107406, Dec. 2022. ISSN 0747-5632. doi: 10.1016/j.chb.2022.107406. URL https://www.sciencedirect.com/science/article/pii/S074756322200228X.

[77] J. Chien, K. R. McKee, J. Kay, and W. Isaac. Recourse for reclamation: Chatting with generative language models, Mar. 2024. URL http://arxiv.org/abs/2403.14467.

[78] J. Cho, A. Zala, and M. Bansal. DALL-Eval: Probing the reasoning skills and social biases of text-to-image generation models. In *ICCV*, 2023.

[79] D. Y. Choi and J. H. Kang. Net Job Creation in an Increasingly Autonomous Economy: The Challenge of a Generation. *Journal of Management Inquiry*, 28(3):300–305, July 2019. ISSN 1056-4926. doi: 10.1177/1056492619827372. URL https://doi.org/10.1177/1056492619827372.

[80] R. Chowdhury. *Technology-Facilitated Gender-Based Violencein an Era of Generative AI*. World Trends in Freedom of Expression and Media Development Series. UNESCO, France, 2023. ISBN 978-92-3-100631-9. URL https://unesdoc.unesco.org/ark:/48223/pf0000387483.

[81] M. Cifor, P. Garcia, T. L. Cowan, J. Rault, J. Sutherland, A. L. Hoffman, N. Salehi, and L. Nakamura. Feminist Data Manifest-No, 2019. URL https://www.manifestno.com.

[82] T. Claburn. AI assistants help developers produce code that's insecure. *The Register*, 2022. URL https://www.theregister.com/2022/12/21/ai_assistants_bad_code/.

[83] CodeCarbon. CodeCarbon. mlco2, May 2023. URL https://github.com/mlco2/codecarbon.

[84] D. Coffey. Māori are trying to save their language from Big Tech. *Wired*, 2021. ISSN 1059-1028. URL https://www.wired.com/story/maori-language-tech/.

[85] N. Couldry and U. A. Mejias. The decolonial turn in data and technology research: What is at stake and where is it heading? *Information, Communication & Society*, 26(4):786–802, Mar. 2023. ISSN 1369-118X. doi: 10.1080/1369118X.2021.1986102. URL https://doi.org/10.1080/1369118X.2021.1986102.

[86] K. Crenshaw. Mapping the Margins: Intersectionality, Identity Politics, and Violence against Women of Color. *Stanford Law Review*, 43(6):1241, July 1991. ISSN 00389765. doi: 10.2307/1229039. URL https://www.jstor.org/stable/1229039?origin=crossref.

[87] B. Dang, M. J. Riedl, and M. Lease. But Who Protects the Moderators? The Case of Crowdsourced Image Moderation, Jan. 2020. URL http://arxiv.org/abs/1804.10999.

[88] N. C. Dang, M. N. Moreno-García, and F. De La Prieta. Sentiment Analysis Based on Deep Learning: A Comparative Study. *Electronics*, 9(3):483, Mar. 2020. ISSN 2079-9292. doi: 10.3390/electronics9030483. URL https://www.mdpi.com/2079-9292/9/3/483.

[89] Databricks. Databricks pricing, Fri, 10/21/2022 - 20:37. URL https://www.databricks.com/product/pricing.

[90] T. Davidson and D. Bhattacharya. Examining Racial Bias in an Online Abuse Corpus with Structural Topic Modeling. In *ICWSM Data Challenge*, May 2020. URL http://arxiv.org/abs/2005.13041.

[91] T. Davidson, D. Bhattacharya, and I. Weber. Racial Bias in Hate Speech and Abusive Language Detection Datasets. In *Proceedings of the Third Workshop on Abusive Language Online*, pages 25–35, Florence, Italy, 2019. Association for Computational Linguistics. doi: 10.18653/v1/W19-3504. URL https://www.aclweb.org/anthology/W19-3504.

[92] Department of Defense. DOD Adopts Ethical Principles for Artificial Intelligence, 2020. URL https://www.defense.gov/News/Releases/Release/Article/2091996/dod-adopts-ethical-principles-for-artificial-intelligence/.

[93] Department of International Cooperation Ministry of Science and Technology. Next Generation Artificial Intelligence Development Plan. Technical report, Department of International Cooperation,Ministry of Science and Technology (MOST), Beijing, China, 2017. URL http://fi.china-embassy.gov.cn/eng/kxjs/201710/P020210628714286134479.pdf.

[94] S. Dev, M. Monajatipoor, A. Ovalle, A. Subramonian, J. Phillips, and K.-W. Chang. Harms of Gender Exclusivity and Challenges in Non-Binary Representation in Language Technologies. In M.-F. Moens, X. Huang, L. Specia, and S. W.-t. Yih, editors, *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1968–1994, Online and Punta Cana, Dominican Republic, Nov. 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.150. URL https://aclanthology.org/2021.emnlp-main.150.

[95] S. Dev, J. Goyal, D. Tewari, S. Dave, and V. Prabhakaran. Building socio-culturally inclusive stereotype resources with community engagement. In A. Oh, T. Neumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 4365–4381. Curran Associates, Inc., 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/0dc91de822b71c66a7f54fa121d8cbb9-Paper-Datasets_and_Benchmarks.pdf.

[96] H. Devinney, J. Björklund, and H. Björklund. Semi-supervised topic modeling for gender bias discovery in English and Swedish. In M. R. Costa-jussà, C. Hardmeier, W. Radford, and K. Webster, editors, *Proceedings of the Second Workshop on Gender Bias in Natural Language Processing*, pages 79–92, Barcelona, Spain (Online), Dec. 2020. Association for Computational Linguistics. URL https://aclanthology.org/2020.gebnlp-1.8.

[97] M. Díaz, I. Kivlichan, R. Rosen, D. Baker, R. Amironesei, V. Prabhakaran, and E. Denton. CrowdWorkSheets: Accounting for Individual and Collective Identities Underlying Crowdsourced Dataset Annotation. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 2342–2351, Seoul Republic of Korea, June 2022. ACM. ISBN 978-1-4503-9352-2. doi: 10.1145/3531146.3534647. URL https://dl.acm.org/doi/10.1145/3531146.3534647.

[98] E. Dinan, G. Abercrombie, A. S. Bergman, S. Spruit, D. Hovy, Y.-L. Boureau, and V. Rieser. Anticipating Safety Issues in E2E Conversational AI: Framework and Tooling, July 2021. URL http://arxiv.org/abs/2107.03451.

[99] J. Dodge, M. Sap, A. Marasović, W. Agnew, G. Ilharco, D. Groeneveld, M. Mitchell, and M. Gardner. Documenting Large Webtext Corpora: A Case Study on the Colossal Clean Crawled Corpus. In M.-F. Moens, X. Huang, L. Specia, and S. W.-t. Yih, editors, *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1286–1305, Online and Punta Cana, Dominican Republic, Nov. 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.98. URL https://aclanthology.org/2021.emnlp-main.98.

[100] M. Douglas. *Purity and Danger: An Analysis of the Concepts of Pollution and Taboo*. Routledge, London, repr edition, 1978. ISBN 978-0-7100-1299-9 978-0-7100-8827-7.

[101] C. Douwes, P. Esling, and J.-P. Briot. Energy Consumption of Deep Generative Audio Models, Oct. 2021. URL http://arxiv.org/abs/2107.02621.

[102] Dreyfus. Generative AI and the protection of intellectual property rights, May 2023. URL https://www.dreyfus.fr/en/2023/05/22/generative-ai-balancing-innovation-and-intellectual-property-rights-protection/.

[103] M. Duan, A. Suri, N. Mireshghallah, S. Min, W. Shi, L. Zettlemoyer, Y. Tsvetkov, Y. Choi, D. Evans, and H. Hajishirzi. Do Membership Inference Attacks Work on Large Language Models?, Feb. 2024. URL http://arxiv.org/abs/2402.07841.

[104] Y. Elazar, A. Bhagia, I. Magnusson, A. Ravichander, D. Schwenk, A. Suhr, P. Walsh, D. Groeneveld, L. Soldaini, S. Singh, H. Hajishirzi, N. A. Smith, and J. Dodge. What's In My Big Data?, Mar. 2024. URL http://arxiv.org/abs/2310.20707.

[105] C. Ess. Ethical pluralism and global information ethics. *Ethics and Information Technology*, 8 (4):215–226, Nov. 2006. ISSN 1572-8439. doi: 10.1007/s10676-006-9113-3. URL https://doi.org/10.1007/s10676-006-9113-3.

[106] European Commission. Article 7 GDPR. Conditions for consent | GDPR-Text.com, 2016. URL https://gdpr-text.com/read/article-7/.

[107] European Commission. Proposal for a Regulation laying down harmonised rules on artificial intelligence | Shaping Europe's digital future, Apr. 2021. URL https://digital-strategy.ec.europa.eu/en/library/proposal-regulation-laying-down-harmonised-rules-artificial-intelligence.

[108] European Institute for Gender Equality. Cyber violence is a growing threat, especially for women and girls, 2017. URL https://eige.europa.eu/newsroom/news/cyber-violence-growing-threat-especially-women-and-girls?language_content_entity=en.

[109] Fairwork. Fairwork. URL https://fair.work/en/fw/homepage/.

[110] H. Farid. Creating, Using, Misusing, and Detecting Deep Fakes. *Journal of Online Trust and Safety*, 1(4), Sept. 2022. ISSN 2770-3142. doi: 10.54501/jots.v1i4.56. URL https://tsjournal.org/index.php/jots/article/view/56.

[111] Federal Trade Commission. Protections Against Discrimination and Other Prohibited Practices, 2003. URL https://www.ftc.gov/policy-notices/no-fear-act/protections-against-discrimination.

[112] Federal Trade Commission. FTC Proposes New Protections to Combat AI Impersonation of Individuals, Feb. 2024. URL https://www.ftc.gov/news-events/news/press-releases/2024/02/ftc-proposes-new-protections-combat-ai-impersonation-individuals.

[113] P. Feldman, A. Dant, and D. Rosenbluth. Ethics, Rules of Engagement, and AI: Neural Narrative Mapping Using Large Transformer Language Models, Feb. 2022. URL http://arxiv.org/abs/2202.02647.

[114] V. Feldman. Does learning require memorization? a short tale about a long tail. In *Proceedings of the 52nd Annual ACM SIGACT Symposium on Theory of Computing*, pages 954–959, Chicago IL USA, June 2020. ACM. ISBN 978-1-4503-6979-4. doi: 10.1145/3357713.3384290. URL https://dl.acm.org/doi/10.1145/3357713.3384290.

[115] D. Feliba. In Latin America, data center plans fuel water worries, 2023. URL https://www.reuters.com/article/idUSL8N3AU1PY/.

[116] G. Fergusson, C. Fitzgerald, C. Frascella, M. Iorio, T. McBrien, C. Schroeder, B. Winters, and E. Zhou. Generating Harms: Generative AI's Impact & Paths Forward. Technical report, Electronic Privacy Information Center, 2023. URL https://epic.org/documents/generating-harms-generative-ais-impact-paths-forward/.

[117] A. Field, S. L. Blodgett, Z. Talat, and Y. Tsvetkov. A Survey of Race, Racism, and Anti-Racism in NLP. 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 1905–1925, Online, 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.149. URL https://aclanthology.org/2021.acl-long.149.

[118] E. E. Fitzsimmons-Craft, W. W. Chan, A. C. Smith, M.-L. Firebaugh, L. A. Fowler, N. Topooco, B. DePietro, D. E. Wilfley, C. B. Taylor, and N. C. Jacobson. Effectiveness of a chatbot for eating disorders prevention: A randomized clinical trial. *International Journal of Eating Disorders*, 55(3):343–353, 2022. ISSN 1098-108X. doi: 10.1002/eat.23662. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/eat.23662.

[119] S. A. Friedler, C. Scheidegger, and S. Venkatasubramanian. The (Im)possibility of fairness: Different value systems require different mechanisms for fair decision making. *Communications of the ACM*, 64(4):136–143, Apr. 2021. ISSN 0001-0782, 1557-7317. doi: 10.1145/3433949. URL https://dl.acm.org/doi/10.1145/3433949.

[120] B. Friedman, P. H. Kahn, A. Borning, and A. Huldtgren. Value Sensitive Design and Information Systems. In N. Doorn, D. Schuurbiers, I. van de Poel, and M. E. Gorman, editors, *Early Engagement and New Technologies: Opening up the Laboratory*, pages 55–95. Springer Netherlands, Dordrecht, 2013. ISBN 978-94-007-7844-3. doi: 10.1007/978-94-007-7844-3_4. URL https://doi.org/10.1007/978-94-007-7844-3_4.

[121] F. Friedrich, M. Brack, L. Struppek, D. Hintersdorf, P. Schramowski, S. Luccioni, and K. Kersting. Fair Diffusion: Instructing Text-to-Image Generation Models on Fairness, July 2023. URL http://arxiv.org/abs/2302.10893.

[122] F. Friedrich, K. Hämmerl, P. Schramowski, J. Libovicky, K. Kersting, and A. Fraser. Multilingual Text-to-Image Generation Magnifies Gender Stereotypes and Prompt Engineering May Not Help You, Jan. 2024. URL http://arxiv.org/abs/2401.16092.

[123] R. Gandikota, J. Materzyńska, J. Fiotto-Kaufman, and D. Bau. Erasing Concepts from Diffusion Models. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2426–2436, Paris, France, Oct. 2023. IEEE. ISBN 9798350307184. doi: 10.1109/ICCV51070.2023.00230. URL https://ieeexplore.ieee.org/document/10378568/.

[124] L. Gao, J. Tow, B. Abbasi, S. Biderman, S. Black, A. DiPofi, C. Foster, L. Golding, J. Hsu, A. Le Noac'h, H. Li, K. McDonell, N. Muennighoff, C. Ociepa, J. Phang, L. Reynolds, H. Schoelkopf, A. Skowron, L. Sutawika, E. Tang, A. Thite, B. Wang, K. Wang, and A. Zou. A framework for few-shot language model evaluation, Dec. 2023. URL https://zenodo.org/records/10256836.

[125] M. Gaske. Corporate Officers' Fiduciary Duty to Monitor Generative Artificial Intelligence, Dec. 2023. URL https://papers.ssrn.com/abstract=4664899.

[126] W. Gaviria Rojas, S. Diamos, K. Kini, D. Kanter, V. Janapa Reddi, and C. Coleman. The dollar street dataset: Images representing the geographic and socioeconomic diversity of the world. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 12979–12990. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/5474d9d43c0519aa176276ff2c1ca528-Paper-Datasets_and_Benchmarks.pdf.

[127] T. Gebru, J. Morgenstern, B. Vecchione, J. W. Vaughan, H. Wallach, H. D. Iii, and K. Crawford. Datasheets for datasets. *Communications of the ACM*, 64(12):86–92, Dec. 2021. ISSN 0001-0782, 1557-7317. doi: 10.1145/3458723. URL https://dl.acm.org/doi/10.1145/3458723.

[128] R. S. Geiger, K. Yu, Y. Yang, M. Dai, J. Qiu, R. Tang, and J. Huang. Garbage in, garbage out?: Do machine learning application papers in social computing report where human-labeled training data comes from? In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 325–336, Barcelona Spain, Jan. 2020. ACM. ISBN 978-1-4503-6936-7. doi: 10.1145/3351095.3372862. URL https://dl.acm.org/doi/10.1145/3351095.3372862.

[129] A. Ghosh, R. Dutt, and C. Wilson. When Fair Ranking Meets Uncertain Inference. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1033–1043, Virtual Event Canada, July 2021. ACM. ISBN 978-1-4503-8037-9. doi: 10.1145/3404835.3462850. URL https://dl.acm.org/doi/10.1145/3404835.3462850.

[130] N. Gillespie. Trust in artificial intelligence - KPMG Global, Jan. 2024. URL https://kpmg.com/xx/en/home/insights/2023/09/trust-in-artificial-intelligence.html.

[131] A. Glaese, N. McAleese, M. Trębacz, J. Aslanides, V. Firoiu, T. Ewalds, M. Rauh, L. Weidinger, M. Chadwick, P. Thacker, L. Campbell-Gillingham, J. Uesato, P.-S. Huang, R. Comanescu, F. Yang, A. See, S. Dathathri, R. Greig, C. Chen, D. Fritz, J. S. Elias, R. Green, S. Mokrá, N. Fernando, B. Wu, R. Foley, S. Young, I. Gabriel, W. Isaac, J. Mellor, D. Hassabis, K. Kavukcuoglu, L. A. Hendricks, and G. Irving. Improving alignment of dialogue agents via targeted human judgements, Sept. 2022. URL http://arxiv.org/abs/2209.14375.

[132] W. Godel, Z. Sanderson, K. Aslett, J. Nagler, R. Bonneau, N. Persily, and J. Tucker. Moderating with the Mob: Evaluating the Efficacy of Real-Time Crowdsourced Fact-Checking. *Journal of Online Trust and Safety*, 1(1), Oct. 2021. ISSN 2770-3142. doi: 10.54501/jots.v1i1.15. URL https://tsjournal.org/index.php/jots/article/view/15.

[133] U. Gohar and L. Cheng. A Survey on Intersectional Fairness in Machine Learning: Notions, Mitigation, and Challenges. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*, pages 6619–6627, Macau, SAR China, Aug. 2023. International Joint Conferences on Artificial Intelligence Organization. ISBN 978-1-956792-03-4. doi: 10.24963/ijcai.2023/742. URL https://www.ijcai.org/proceedings/2023/742.

[134] U. Gohar, S. Biswas, and H. Rajan. Towards understanding fairness and its composition in ensemble machine learning. In *2023 IEEE/ACM 45th International Conference on Software Engineering (ICSE)*, pages 1533–1545. IEEE, 2023.

[135] U. Gohar, M. C. Hunter, A. Marczak-Czajka, R. R. Lutz, M. B. Cohen, and J. Cleland-Huang. Towards engineering fair and equitable software systems for managing low-altitude airspace authorizations. *arXiv preprint arXiv:2401.07353*, 2024.

[136] S. Goldfarb-Tarrant, R. Marchant, R. Muñoz Sánchez, M. Pandya, and A. Lopez. Intrinsic Bias Metrics Do Not Correlate with Application Bias. 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 1926–1940, Online, 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.150. URL https://aclanthology.org/2021.acl-long.150.

[137] S. Goldfarb-Tarrant, E. Ungless, E. Balkir, and S. L. Blodgett. This prompt is measuring <mask>: Evaluating bias evaluation in language models. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 2209–2225, Toronto, Canada, 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-acl.139. URL https://aclanthology.org/2023.findings-acl.139.

[138] J. A. Goldstein, G. Sastry, M. Musser, R. DiResta, M. Gentzel, and K. Sedova. Generative Language Models and Automated Influence Operations: Emerging Threats and Potential Mitigations, Jan. 2023. URL http://arxiv.org/abs/2301.04246.

[139] J. A. Goldstein, J. Chao, S. Grossman, A. Stamos, and M. Tomz. How persuasive is AI-generated propaganda? *PNAS Nexus*, 3(2):pgae034, Feb. 2024. ISSN 2752-6542. doi: 10.1093/pnasnexus/pgae034. URL https://academic.oup.com/pnasnexus/article/doi/10.1093/pnasnexus/pgae034/7610937.

[140] Google Cloud. Pricing | Vertex AI | Google Cloud. URL https://cloud.google.com/vertex-ai/pricing.

[141] Y. Green, A. Gully, Y. Roth, A. Roy, J. A. Tucker, and A. Wanless. Evidence-Based Misinformation Interventions: Challenges and Opportunities for Measurement and Collaboration, 2023. URL https://carnegieendowment.org/2023/01/09/evidence-based-misinformation-interventions-challenges-and-opportunities-for-measurement-and-collaboration-pub-88661.

[142] X. Gu, C. Du, T. Pang, C. Li, M. Lin, and Y. Wang. On Memorization in Diffusion Models, Oct. 2023. URL http://arxiv.org/abs/2310.02664.

[143] A. M. Guess, P. Barberá, S. Munzert, and J. Yang. The consequences of online partisan media. *Proceedings of the National Academy of Sciences*, 118(14):e2013464118, Apr. 2021. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.2013464118. URL https://pnas.org/doi/full/10.1073/pnas.2013464118.

[144] U. Gupta, Y. G. Kim, S. Lee, J. Tse, H.-H. S. Lee, G.-Y. Wei, D. Brooks, and C.-J. Wu. Chasing Carbon: The Elusive Environmental Footprint of Computing. *IEEE Micro*, 42(4): 37–47, July 2022. ISSN 0272-1732, 1937-4143. doi: 10.1109/MM.2022.3163226. URL https://ieeexplore.ieee.org/document/9744492/.

[145] C. Haerpfer, R. Inglehart, A. Moreno, C. Welzel, K. Kizilova, J. Diez-Medrano, M. Lagos, P. Norris, E. Ponarin, and B. Puranen. World Values Survey Wave 7 (2017-2022) Cross-National Data-Set, 2022. URL http://www.worldvaluessurvey.org/WVSDocumentationWV7.jsp.

[146] O. L. Haimson, D. Delmonaco, P. Nie, and A. Wegner. Disproportionate Removals and Differing Content Moderation Experiences for Conservative, Transgender, and Black Social Media Users: Marginalization and Moderation Gray Areas. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW2):466:1–466:35, Oct. 2021. doi: 10.1145/3479610. URL https://dl.acm.org/doi/10.1145/3479610.

[147] T. Hartvigsen, S. Gabriel, H. Palangi, M. Sap, D. Ray, and E. Kamar. ToxiGen: A Large-Scale Machine-Generated Dataset for Adversarial and Implicit Hate Speech Detection. In S. Muresan, P. Nakov, and A. Villavicencio, editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3309–3326, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.234. URL https://aclanthology.org/2022.acl-long.234.

[148] W. D. Heaven. Why Meta's latest large language model survived only three days online. *MIT Technology Review*, 2022. URL https://www.technologyreview.com/2022/11/18/1063487/meta-large-language-model-ai-only-survived-three-days-gpt-3-science/.

[149] P. Henderson, J. Hu, J. Romoff, E. Brunskill, D. Jurafsky, and J. Pineau. Towards the Systematic Reporting of the Energy and Carbon Footprints of Machine Learning. *Journal of Machine Learning Research*, 21(248):1–43, 2020. ISSN 1533-7928. URL http://jmlr.org/papers/v21/20-312.html.

[150] D. Hintersdorf, L. Struppek, D. Neider, and K. Kersting. Defending Our Privacy With Backdoors, Feb. 2024. URL http://arxiv.org/abs/2310.08320.

[151] M. Hirsh. How AI Will Revolutionize Warfare, 2023. URL https://foreignpolicy.com/2023/04/11/ai-arms-race-artificial-intelligence-chatgpt-military-technology/.

[152] L. Hofeditz, M. Mirbabaie, J. Holstein, and S. Stieglitz. Do You trust an AI-Journalist? A Credibility Analysis of News Content with AI-Authorship. *ECIS 2021 Research Papers*, June 2021. URL https://aisel.aisnet.org/ecis2021_rp/50.

[153] G. Hofstede. Dimensionalizing Cultures: The Hofstede Model in Context. *Online Readings in Psychology and Culture*, 2(1), Dec. 2011. ISSN 2307-0919. doi: 10.9707/2307-0919.1014. URL https://scholarworks.gvsu.edu/orpc/vol2/iss1/8.

[154] G. Hofstede. *Culture's Consequences: Comparing Values, Behaviors, Institutions, and Organizations across Nations*. Sage, Thousand Oaks, Calif., 2. ed. [nachdr.] edition, 2013. ISBN 978-0-8039-7324-4 978-0-8039-7323-7.

[155] Home Security Heroes. STATE OF DEEPFAKES: Realities, Threats, and Impact., 2023. URL https://blog.biocomm.ai/2024/02/18/report-2023-state-of-deepfakes-realities-threats-and-impact/.

[156] S. Hooker. Moving beyond "algorithmic bias is a data problem". *Patterns*, 2(4): 100241, Apr. 2021. ISSN 26663899. doi: 10.1016/j.patter.2021.100241. URL https://linkinghub.elsevier.com/retrieve/pii/S2666389921000611.

[157] S. Hooker, N. Moorosi, G. Clark, S. Bengio, and E. Denton. Characterising Bias in Compressed Models, Dec. 2020. URL http://arxiv.org/abs/2010.03058.

[158] House of Commons. Digital Charter Implementation Act, 2022. URL https://www.parl.ca/DocumentViewer/en/44-1/bill/C-27/first-reading.

[159] D. Hovy and S. L. Spruit. The Social Impact of Natural Language Processing. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 591–598, Berlin, Germany, 2016. Association for Computational Linguistics. doi: 10.18653/v1/P16-2096. URL http://aclweb.org/anthology/P16-2096.

[160] H. Huang, T. Tang, D. Zhang, X. Zhao, T. Song, Y. Xia, and F. Wei. Not All Languages Are Created Equal in LLMs: Improving Multilingual Capability by Cross-Lingual-Thought Prompting. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 12365–12394, Singapore, 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.826. URL https://aclanthology.org/2023.findings-emnlp.826.

[161] Human Rights Watch. Agenda for action: Alternative processes for negotiating a killer robots treaty, 2022. URL https://www.hrw.org/report/2022/11/10/agenda-action/alternative-processes-negotiating-killer-robots-treaty.

[162] B. Hutchinson and M. Mitchell. 50 Years of Test (Un)fairness: Lessons for Machine Learning. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 49–58, Atlanta GA USA, Jan. 2019. ACM. ISBN 978-1-4503-6125-5. doi: 10.1145/3287560.3287600. URL https://dl.acm.org/doi/10.1145/3287560.3287600.

[163] B. Hutchinson, N. Rostamzadeh, C. Greer, K. Heller, and V. Prabhakaran. Evaluation Gaps in Machine Learning Practice. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 1859–1876, Seoul Republic of Korea, June 2022. ACM. ISBN 978-1-4503-9352-2. doi: 10.1145/3531146.3533233. URL https://dl.acm.org/doi/10.1145/3531146.3533233.

[164] A. N. Institute. Zero Trust AI Governance, Aug. 2023. URL https://ainowinstitute.org/publication/zero-trust-ai-governance.

[165] Institute of Medicine (US) Committee on Understanding and Eliminating Racial and Ethnic Disparities in Health Care. *Unequal Treatment: Confronting Racial and Ethnic Disparities in Health Care*. National Academies Press (US), Washington (DC), 2003. URL http://www.ncbi.nlm.nih.gov/books/NBK220358/. Eds. Smedley, Brian D. and Stith, Adrienne Y. and Nelson, Alan R.

[166] Y. Ishibashi and H. Shimodaira. Knowledge Sanitization of Large Language Models, Mar. 2024. URL http://arxiv.org/abs/2309.11852.

[167] J. James, V. Yogarajan, I. Shields, C. Watson, P. Keegan, K. Mahelona, and P.-L. Jones. Language Models for Code-switch Detection of te reo Māori and English in a Low-resource Setting. In M. Carpuat, M.-C. de Marneffe, and I. V. Meza Ruiz, editors, *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 650–660, Seattle, United States, July 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-naacl.49. URL https://aclanthology.org/2022.findings-naacl.49.

[168] Y. Jernite. Let's talk about biases in machine learning! Ethics and Society Newsletter #2, 2022. URL https://huggingface.co/blog/ethics-soc-2.

[169] Y. Jernite, H. Nguyen, S. Biderman, A. Rogers, M. Masoud, V. Danchev, S. Tan, A. S. Luccioni, N. Subramani, I. Johnson, G. Dupont, J. Dodge, K. Lo, Z. Talat, D. Radev, A. Gokaslan, S. Nikpoor, P. Henderson, R. Bommasani, and M. Mitchell. Data Governance in the Age of Large-Scale Data-Driven Language Technology. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 2206–2222, Seoul Republic of Korea, June 2022. ACM. ISBN 978-1-4503-9352-2. doi: 10.1145/3531146.3534637. URL https://dl.acm.org/doi/10.1145/3531146.3534637.

[170] Z. Ji, N. Lee, R. Frieske, T. Yu, D. Su, Y. Xu, E. Ishii, Y. Bang, D. Chen, H. S. Chan, W. Dai, A. Madotto, and P. Fung. Survey of Hallucination in Natural Language Generation. *ACM Computing Surveys*, 55(12):1–38, Dec. 2023. ISSN 0360-0300, 1557-7341. doi: 10.1145/3571730. URL http://arxiv.org/abs/2202.03629.

[171] H. Jiang, D. Beeferman, B. Roy, and D. Roy. CommunityLM: Probing partisan worldviews from language models. In N. Calzolari, C.-R. Huang, H. Kim, J. Pustejovsky, L. Wanner, K.-S. Choi, P.-M. Ryu, H.-H. Chen, L. Donatelli, H. Ji, S. Kurohashi, P. Paggio, N. Xue, S. Kim, Y. Hahm, Z. He, T. K. Lee, E. Santus, F. Bond, and S.-H. Na, editors, *Proceedings of the 29th International Conference on Computational Linguistics*, pages 6818–6826, Gyeongju, Republic of Korea, Oct. 2022. International Committee on Computational Linguistics. URL https://aclanthology.org/2022.coling-1.593.

[172] H. H. Jiang, L. Brown, J. Cheng, M. Khan, A. Gupta, D. Workman, A. Hanna, J. Flowers, and T. Gebru. AI Art and its Impact on Artists. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '23, pages 363–374, New York, NY, USA, Aug. 2023. Association for Computing Machinery. ISBN 9798400702310. doi: 10.1145/3600211.3604681. URL https://dl.acm.org/doi/10.1145/3600211.3604681.

[173] S. Jindal. Responsible Sourcing of Data Enrichment Services, June 2021. URL https://partnershiponai.org/responsible-sourcing-considerations/.

[174] P. Joshi, S. Santy, A. Budhiraja, K. Bali, and M. Choudhury. The State and Fate of Linguistic Diversity and Inclusion in the NLP World. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online, 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.560. URL https://www.aclweb.org/anthology/2020.acl-main.560.

[175] L. H. Kaack, P. L. Donti, E. Strubell, G. Kamiya, F. Creutzig, and D. Rolnick. Aligning artificial intelligence with climate change mitigation. *Nature Climate Change*, 12(6):518–527, June 2022. ISSN 1758-678X, 1758-6798. doi: 10.1038/s41558-022-01377-7. URL https://www.nature.com/articles/s41558-022-01377-7.

[176] P. Kalluri. Don't ask if artificial intelligence is good or fair, ask how it shifts power. *Nature*, 583 (7815):169–169, July 2020. ISSN 0028-0836, 1476-4687. doi: 10.1038/d41586-020-02003-2. URL http://www.nature.com/articles/d41586-020-02003-2.

[177] M. Kamal and W. J. Newman. Revenge Pornography: Mental Health Implications and Related Legislation. *The Journal of the American Academy of Psychiatry and the Law*, 44(3):359–367, Sept. 2016. ISSN 1943-3662.

[178] R. Kamikubo, U. Dwivedi, and H. Kacorri. Sharing Practices for Datasets Related to Accessibility and Aging. In *Proceedings of the 23rd International ACM SIGACCESS Conference on Computers and Accessibility*, pages 1–16, Virtual Event USA, Oct. 2021. ACM. ISBN 978-1-4503-8306-6. doi: 10.1145/3441852.3471208. URL https://dl.acm.org/doi/10.1145/3441852.3471208.

[179] M. E. Kaminski. Regulating the Risks of AI. *Boston University Law Review*, 103(1347), 2022. ISSN 1556-5068. doi: 10.2139/ssrn.4195066. URL https://www.ssrn.com/abstract=4195066.

[180] M. Kan. US Lawmakers Push FTC to Crack Down on VPNs That Use Deceptive Practices. *PC Mag*, 2022. URL https://www.pcmag.com/news/us-lawmakers-push-ftc-to-crack-down-on-vpns-that-use-deceptive-practices.

[181] F. Karimi. 'Mom, these bad men have me': She believes scammers cloned her daughter's voice in a fake kidnapping, Apr. 2023. URL https://www.cnn.com/2023/04/29/us/ai-scam-calls-kidnapping-cec/index.html.

[182] J. Kaur, R. A. Dara, C. Obimbo, F. Song, and K. Menard. A comprehensive keyword analysis of online privacy policies. *Information Security Journal: A Global Perspective*, 27(5-6): 260–275, Nov. 2018. ISSN 1939-3555, 1939-3547. doi: 10.1080/19393555.2019.1606368. URL https://www.tandfonline.com/doi/full/10.1080/19393555.2019.1606368.

[183] D. Kaushik, Z. C. Lipton, and A. J. London. Resolving the Human-subjects Status of Machine Learning's Crowdworkers: What ethical framework should govern the interaction of ML researchers and crowdworkers? *Queue*, 21(6):101–127, Dec. 2023. ISSN 1542-7730, 1542-7749. doi: 10.1145/3639452. URL https://dl.acm.org/doi/10.1145/3639452.

[184] K. Kelley. Good practice in the conduct and reporting of survey research. *International Journal for Quality in Health Care*, 15(3):261–266, May 2003. ISSN 13534505, 14643677. doi: 10.1093/intqhc/mzg031. URL https://academic.oup.com/intqhc/article-lookup/doi/10.1093/intqhc/mzg031.

[185] Z. Khan and Y. Fu. One Label, One Billion Faces: Usage and Consistency of Racial Categories in Computer Vision. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 587–597, Virtual Event Canada, Mar. 2021. ACM. ISBN 978-1-4503-8309-7. doi: 10.1145/3442188.3445920. URL https://dl.acm.org/doi/10.1145/3442188.3445920.

[186] H. Khlaaf, P. Mishkin, J. Achiam, G. Krueger, and M. Brundage. A Hazard Analysis Framework for Code Synthesis Large Language Models, July 2022. URL http://arxiv.org/abs/2207.14157.

[187] D. Kiela, M. Bartolo, Y. Nie, D. Kaushik, A. Geiger, Z. Wu, B. Vidgen, G. Prasad, A. Singh, P. Ringshia, Z. Ma, T. Thrush, S. Riedel, Z. Talat, P. Stenetorp, R. Jia, M. Bansal, C. Potts, and A. Williams. Dynabench: Rethinking Benchmarking in NLP. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4110–4124, Online, 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.324. URL https://www.aclweb.org/anthology/2021.naacl-main.324.

[188] S. Kim, S. Yun, H. Lee, M. Gubri, S. Yoon, and S. J. Oh. ProPILE: Probing privacy leakage in large language models. In A. Oh, T. Neumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 20750–20762. Curran Associates, Inc., 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/420678bb4c8251ab30e765bc27c3b047-Paper-Conference.pdf.

[189] J. King and C. Meinhardt. Rethinking Privacy in the AI Era: Policy Provocations for a Data-Centric World. Technical report, Stanford, 2024. URL https://hai.stanford.edu/white-paper-rethinking-privacy-ai-era-policy-provocations-data-centric-world.

[190] J. Kirchenbauer, J. Geiping, Y. Wen, J. Katz, I. Miers, and T. Goldstein. A Watermark for Large Language Models, June 2023. URL http://arxiv.org/abs/2301.10226.

[191] K. Klyman. Acceptable use policies for foundation models: Considerations for policymakers and developers, Apr. 2024. URL https://crfm.stanford.edu/2024/04/08/aups.html.

[192] W.-Y. Ko, D. D'souza, K. Nguyen, R. Balestriero, and S. Hooker. FAIR-Ensemble: When Fairness Naturally Emerges From Deep Ensembling, Dec. 2023. URL http://arxiv.org/abs/2303.00586.

[193] A. Korinek and J. Vipra. Market concentration implications of foundation models: The Invisible Hand of ChatGPT. *Brookings Institution*, Sept. 2023. URL https://policycommons.net/artifacts/4864844/market-concentration-implications-of-foundation-models/5702045/.

[194] S. M. Labott, T. P. Johnson, M. Fendrich, and N. C. Feeny. Emotional Risks to Respondents in Survey Research: Some Empirical Evidence. *Journal of Empirical Research on Human Research Ethics*, 8(4):53–66, Oct. 2013. ISSN 1556-2646, 1556-2654. doi: 10.1525/jer.2013.8.4.53. URL http://journals.sagepub.com/doi/10.1525/jer.2013.8.4.53.

[195] A. Lacoste, A. Luccioni, V. Schmidt, and T. Dandres. Quantifying the Carbon Emissions of Machine Learning, Nov. 2019. URL http://arxiv.org/abs/1910.09700.

[196] J. Lalor, Y. Yang, K. Smith, N. Forsgren, and A. Abbasi. Benchmarking Intersectional Biases in NLP. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3598–3609, Seattle, United States, 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.naacl-main.263. URL https://aclanthology.org/2022.naacl-main.263.

[197] J. Larson, S. Mattu, L. Kirchner, and J. Angwin. How We Analyzed the COMPAS Recidivism Algorithm, 2016. URL https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm.

[198] J.-U. Lee, H. Puerto, B. van Aken, Y. Arase, J. Z. Forde, L. Derczynski, A. Rücklé, I. Gurevych, R. Schwartz, E. Strubell, and J. Dodge. Surveying (Dis)Parities and Concerns of Compute Hungry NLP Research, Nov. 2023. URL http://arxiv.org/abs/2306.16900.

[199] K. Lee, D. Ippolito, A. Nystrom, C. Zhang, D. Eck, C. Callison-Burch, and N. Carlini. Deduplicating Training Data Makes Language Models Better. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8424–8445, Dublin, Ireland, 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.577. URL https://aclanthology.org/2022.acl-long.577.

[200] J. Lepawsky. No insides on the outsides. *Discard Studies*, 0(0), Sept. 2019. URL https://discardstudies.com/2019/09/23/no-insides-on-the-outsides/.

[201] D. I. Levine. Automation as Part of the Solution. *Journal of Management Inquiry*, 28(3): 316–318, July 2019. ISSN 1056-4926. doi: 10.1177/1056492619827375. URL https://doi.org/10.1177/1056492619827375.

[202] B. Li, S. Haider, and C. Callison-Burch. This Land is {Your, My} Land: Evaluating Geopolitical Biases in Language Models, Apr. 2024. URL http://arxiv.org/abs/2305.14610.

[203] C. Li. OpenAI's GPT-3 Language Model: A Technical Overview, June 2020. URL https://lambdalabs.com/blog/demystifying-gpt-3.

[204] H. Li, D. Guo, W. Fan, M. Xu, J. Huang, F. Meng, and Y. Song. Multi-step Jailbreaking Privacy Attacks on ChatGPT. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 4138–4153, Singapore, 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.272. URL https://aclanthology.org/2023.findings-emnlp.272.

[205] N. Li, A. Pan, A. Gopal, S. Yue, D. Berrios, A. Gatti, J. D. Li, A.-K. Dombrowski, S. Goel, L. Phan, G. Mukobi, N. Helm-Burger, R. Lababidi, L. Justen, A. B. Liu, M. Chen, I. Barrass, O. Zhang, X. Zhu, R. Tamirisa, B. Bharathi, A. Khoja, Z. Zhao, A. Herbert-Voss, C. B. Breuer, A. Zou, M. Mazeika, Z. Wang, P. Oswal, W. Liu, A. A. Hunt, J. Tienken-Harder, K. Y. Shih, K. Talley, J. Guan, R. Kaplan, I. Steneker, D. Campbell, B. Jokubaitis, A. Levinson, J. Wang, W. Qian, K. K. Karmakar, S. Basart, S. Fitz, M. Levine, P. Kumaraguru, U. Tupakula, V. Varadharajan, Y. Shoshitaishvili, J. Ba, K. M. Esvelt, A. Wang, and D. Hendrycks. The WMDP Benchmark: Measuring and Reducing Malicious Use With Unlearning, Mar. 2024. URL http://arxiv.org/abs/2403.03218.

[206] T. Li, D. Khashabi, T. Khot, A. Sabharwal, and V. Srikumar. UNQOVERing Stereotyping Biases via Underspecified Questions. In T. Cohn, Y. He, and Y. Liu, editors, *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3475–3489, Online, Nov. 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.findings-emnlp.311. URL https://aclanthology.org/2020.findings-emnlp.311.

[207] W. Liang, M. Yuksekgonul, Y. Mao, E. Wu, and J. Zou. GPT detectors are biased against non-native English writers. *Patterns*, 4(7):100779, July 2023. ISSN 26663899. doi: 10.1016/j.patter.2023.100779. URL https://linkinghub.elsevier.com/retrieve/pii/S2666389923001307.

[208] Q. V. Liao and Z. Xiao. Rethinking Model Evaluation as Narrowing the Socio-Technical Gap, June 2023. URL http://arxiv.org/abs/2306.03100.

[209] M. Liboiron. *Pollution Is Colonialism*. Duke University Press, Durham, 2021. ISBN 978-1-4780-2144-5.

[210] B. Liu, L. Wang, C. Lyu, Y. Zhang, J. Su, S. Shi, and Z. Tu. On the Cultural Gap in Text-to-Image Generation, July 2023. URL http://arxiv.org/abs/2307.02971.

[211] G. Liu, X. Wang, L. Yuan, Y. Chen, and H. Peng. Examining LLMs' Uncertainty Expression Towards Questions Outside Parametric Knowledge, Feb. 2024. URL http://arxiv.org/abs/2311.09731.

[212] L. T. Liu, S. Dean, E. Rolf, M. Simchowitz, and M. Hardt. Delayed Impact of Fair Machine Learning. *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence*, pages 6196–6200, 2019. URL https://www.ijcai.org/proceedings/2019/862.

[213] Z. Liu, U. Iqbal, and N. Saxena. Opted Out, Yet Tracked: Are Regulations Enough to Protect Your Privacy? *Proceedings on Privacy Enhancing Technologies*, 2024. ISSN 2299-0984. URL https://petsymposium.org/popets/2024/popets-2024-0016.php.

[214] Z. Liu, P. Schaldenbrand, B.-C. Okogwu, W. Peng, Y. Yun, A. Hundt, J. Kim, and J. Oh. SCoFT: Self-Contrastive Fine-Tuning for Equitable Image Generation, Jan. 2024. URL http://arxiv.org/abs/2401.08053.

[215] A. S. Luccioni, C. Akiki, M. Mitchell, and Y. Jernite. Stable Bias: Analyzing Societal Representations in Diffusion Models, Nov. 2023. URL http://arxiv.org/abs/2303.11408.

[216] A. S. Luccioni, S. Viguier, and A.-L. Ligozat. Estimating the Carbon Footprint of BLOOM, a 176B Parameter Language Model. *Journal of Machine Learning Research*, 24(253):1–15, 2023. ISSN 1533-7928. URL http://jmlr.org/papers/v24/23-0069.html.

[217] S. Luccioni, Y. Jernite, and M. Mitchell. Introducing the Data Measurements Tool: An Interactive Tool for Looking at Datasets, 2021. URL https://huggingface.co/blog/data-measurements-tool.

[218] S. Luccioni, Y. Jernite, and E. Strubell. Power hungry processing: Watts driving the cost of ai deployment? In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '24, page 85–99, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400704505. doi: 10.1145/3630106.3658542. URL https://doi.org/10.1145/3630106.3658542.

[219] V. Malik, S. Dev, A. Nishi, N. Peng, and K.-W. Chang. Socially Aware Bias Measurements for Hindi Language Representations. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1041–1052, Seattle, United States, 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.naacl-main.76. URL https://aclanthology.org/2022.naacl-main.76.

[220] G. Marcus and R. Southen. Generative AI Has a Visual Plagiarism Problem, 2024. URL https://spectrum.ieee.org/midjourney-copyright.

[221] V. Marian. Opinion | AI could cause a mass-extinction of languages — and ways of thinking. *Washington Post*, Apr. 2023. ISSN 0190-8286. URL https://www.washingtonpost.com/opinions/2023/04/19/ai-chatgpt-language-extinction/.

[222] A. Markdalen, M. Roberts, B. Schwartz, S. KVJ, M. Oost, S. Jones, L. Mitnick, S. Cherian, R. Engels, R. Tolido, M. A. Sowa, R. Puttur, V. Perhirin, S. Andersson, and J. Buvat. Creative and generative AI, Dec. 2023. URL https://www.capgemini.com/insights/research-library/creative-and-generative-ai/.

[223] E. Markey. Markey, Heinrich, Eshoo, Beyer Introduce Legislation to Investigate, Measure Environmental Impacts of Artificial Intelligence, 2024. URL https://www.markey.senate.gov/news/press-releases/markey-heinrich-eshoo-beyer-introduce-legislation-to-investigate-measure-environmental-impacts-of-artificial-intelligence.

[224] K. Martin. The penalty for privacy violations: How privacy violations impact trust online. *Journal of Business Research*, 82:103–116, Jan. 2018. ISSN 0148-2963. doi: 10.1016/j.jbusres.2017.08.034. URL https://www.sciencedirect.com/science/article/pii/S0148296317302965.

[225] C. May, A. Wang, S. Bordia, S. R. Bowman, and R. Rudinger. On Measuring Social Biases in Sentence Encoders. In *Proceedings of the 2019 Conference of the North*, pages 622–628, Minneapolis, Minnesota, 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1063. URL http://aclweb.org/anthology/N19-1063.

[226] J. Mendelsohn, R. Le Bras, Y. Choi, and M. Sap. From Dogwhistles to Bullhorns: Unveiling Coded Rhetoric with Language Models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15162–15180, Toronto, Canada, 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.845. URL https://aclanthology.org/2023.acl-long.845.

[227] M. Miceli, M. Schuessler, and T. Yang. Between Subjectivity and Imposition: Power Dynamics in Data Annotation for Computer Vision. *Proceedings of the ACM on Human-Computer Interaction*, 4(CSCW2):1–25, Oct. 2020. ISSN 2573-0142. doi: 10.1145/3415186. URL https://dl.acm.org/doi/10.1145/3415186.

[228] J. Mickel. Racial/Ethnic Categories in AI and Algorithmic Fairness: Why They Matter and What They Represent, Apr. 2024. URL http://arxiv.org/abs/2404.06717.

[229] Microsoft Threat Intelligence. Staying ahead of threat actors in the age of AI, Feb. 2024. URL https://www.microsoft.com/en-us/security/blog/2024/02/14/staying-ahead-of-threat-actors-in-the-age-of-ai/.

[230] Ministry of Foreign Affairs. REAIM 2023, 2023. URL https://www.government.nl/ministries/ministry-of-foreign-affairs/activiteiten/reaim.

[231] N. Mireshghallah, H. Kim, X. Zhou, Y. Tsvetkov, M. Sap, R. Shokri, and Y. Choi. Can LLMs Keep a Secret? Testing Privacy Implications of Language Models via Contextual Integrity Theory. In *The Twelfth International Conference on Learning Representations*, Oct. 2023. URL https://openreview.net/forum?id=gmg7t8b4s0.

[232] R. Modhvadia. How do people feel about AI? Technical report, Ada Lovelace Institute, 2023. URL https://www.adalovelaceinstitute.org/report/public-attitudes-ai/.

[233] S. Mohamed, M.-T. Png, and W. Isaac. Decolonial AI: Decolonial Theory as Sociotechnical Foresight in Artificial Intelligence. *Philosophy & Technology*, 33(4):659–684, Dec. 2020. ISSN 2210-5433, 2210-5441. doi: 10.1007/s13347-020-00405-8. URL https://link.springer.com/10.1007/s13347-020-00405-8.

[234] I. Mollas, Z. Chrysopoulou, S. Karlos, and G. Tsoumakas. ETHOS: An Online Hate Speech Detection Dataset. *Complex & Intelligent Systems*, 8(6):4663–4678, Dec. 2022. ISSN 2199-4536, 2198-6053. doi: 10.1007/s40747-021-00608-2. URL http://arxiv.org/abs/2006.08328.

[235] A. Monea. *The Digital Closet: How the Internet Became Straight*. The MIT Press, Apr. 2022. ISBN 978-0-262-36913-8. doi: 10.7551/mitpress/12551.001.0001. URL https://direct.mit.edu/books/oa-monograph/5305/The-Digital-ClosetHow-the-Internet-Became-Straight.

[236] S. G. Monserrate. The Cloud Is Material: On the Environmental Impacts of Computation and Data Storage. *MIT Case Studies in Social and Ethical Responsibilities of Computing*, (Winter 2022), Jan. 2022. doi: 10.21428/2c646de5.031d4553. URL https://mit-serc.pubpub.org/pub/the-cloud-is-material/release/2.

[237] B. Moore. *Privacy: Studies in Social and Cultural History*. Routledge Revivals. Routledge, Abingdon, 2018. ISBN 978-1-351-69676-0.

[238] C. Morris. AI chatbot will replace human helpline workers at National Eating Disorder Association. *Fortune Well*, 2023. URL https://fortune.com/well/2023/05/26/national-eating-disorder-association-ai-chatbot-tessa/.

[239] M. Muro and S. Liu. The geography of AI. Technical report, Brookings, 2021. URL https://www.brookings.edu/articles/the-geography-of-ai/.

[240] M. Murphy. Predators Exploit AI Tools to Generate Images of Child Abuse. *Bloomberg.com*, May 2023. URL https://www.bloomberg.com/news/articles/2023-05-23/predators-exploit-ai-tools-to-depict-abuse-prompting-warnings.

[241] M. Nadeem, A. Bethke, and S. Reddy. StereoSet: Measuring stereotypical bias in pretrained language models, Apr. 2020. URL http://arxiv.org/abs/2004.09456.

[242] R. Naik and B. Nushi. Social Biases through the Text-to-Image Generation Lens. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, pages 786–808, Montr\'{e}al QC Canada, Aug. 2023. ACM. ISBN 9798400702310. doi: 10.1145/3600211.3604711. URL https://dl.acm.org/doi/10.1145/3600211.3604711.

[243] N. Nangia, C. Vania, R. Bhalerao, and S. R. Bowman. CrowS-Pairs: A Challenge Dataset for Measuring Social Biases in Masked Language Models. In B. Webber, T. Cohn, Y. He, and Y. Liu, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1953–1967, Online, Nov. 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.154. URL https://aclanthology.org/2020.emnlp-main.154.

[244] M. Nasr, N. Carlini, J. Hayase, M. Jagielski, A. F. Cooper, D. Ippolito, C. A. Choquette-Choo, E. Wallace, F. Tramèr, and K. Lee. Scalable Extraction of Training Data from (Production) Language Models, Nov. 2023. URL http://arxiv.org/abs/2311.17035.

[245] National Institute of Standards and Technology. AI Risk Management Framework: AI RMF (1.0). Technical Report error: NIST AI 100-1, National Institute of Standards and Technology, Gaithersburg, MD, 2023. URL https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-1.pdf.

[246] Neural Information Processing Systems. NeurIPS Code of Ethics, 2023. URL https://neurips.cc/Conferences/2023/EthicsGuidelines.

[247] J. Newman. A Taxonomy of Trustworthiness for Artificial Intelligence. Technical report, Center for Long-Term Cybersecurity, Berkeley, 2023. URL https://cltc.berkeley.edu/publication/a-taxonomy-of-trustworthiness-for-artificial-intelligence/.

[248] D. Nikolaiev. Behind the Millions: Estimating the Scale of Large Language Models, June 2023. URL https://towardsdatascience.com/behind-the-millions-estimating-the-scale-of-large-language-models-97bd7287fb6b.

[249] H. Nissenbaum. *Privacy in Context: Technology, Policy, and the Integrity of Social Life*. Stanford University Press, Stanford, 2009. ISBN 978-0-8047-5236-7.

[250] J. Niu, W. Tang, F. Xu, X. Zhou, and Y. Song. Global Research on Artificial Intelligence from 1990–2014: Spatially-Explicit Bibliometric Analysis. *ISPRS International Journal of Geo-Information*, 5(5):66, May 2016. ISSN 2220-9964. doi: 10.3390/ijgi5050066. URL https://www.mdpi.com/2220-9964/5/5/66.

[251] S. Noy and W. Zhang. Experimental evidence on the productivity effects of generative artificial intelligence. *Science*, 381(6654):187–192, July 2023. doi: 10.1126/science.adh2586. URL https://www.science.org/doi/10.1126/science.adh2586.

[252] D. Nozza, F. Bianchi, and D. Hovy. HONEST: Measuring Hurtful Sentence Completion in Language Models. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2398–2406, Online, 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.191. URL https://aclanthology.org/2021.naacl-main.191.

[253] OECD Policy Observatory. OECD's live repository of AI strategies & policies, 2021. URL https://oecd.ai/en/dashboards.

[254] OECD Policy Observatory. OECD Framework for the Classification of AI Systems: A tool for effective AI policies, Apr. 2023. URL https://oecd.ai/en/classification.

[255] K. Ogueji, O. Ahia, G. Onilude, S. Gehrmann, S. Hooker, and J. Kreutzer. Intriguing Properties of Compression on Multilingual Models. In Y. Goldberg, Z. Kozareva, and Y. Zhang, editors, *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9092–9110, Abu Dhabi, United Arab Emirates, Dec. 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.619. URL https://aclanthology.org/2022.emnlp-main.619.

[256] V. Ooi and G. Goh. Taxation of Automation and Artificial Intelligence as a Tool of Labour Policy. *SSRN Electronic Journal*, 2018. ISSN 1556-5068. doi: 10.2139/ssrn.3322306. URL https://www.ssrn.com/abstract=3322306.

[257] OpenAI. Pricing. URL https://openai.com/pricing.

[258] OpenAI, J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat, R. Avila, I. Babuschkin, S. Balaji, V. Balcom, P. Baltescu, H. Bao, M. Bavarian, J. Belgum, I. Bello, J. Berdine, G. Bernadett-Shapiro, C. Berner, L. Bogdonoff, O. Boiko, M. Boyd, A.-L. Brakman, G. Brockman, T. Brooks, M. Brundage, K. Button, T. Cai, R. Campbell, A. Cann, B. Carey, C. Carlson, R. Carmichael, B. Chan, C. Chang, F. Chantzis, D. Chen, S. Chen, R. Chen, J. Chen, M. Chen, B. Chess, C. Cho, C. Chu, H. W. Chung, D. Cummings, J. Currier, Y. Dai, C. Decareaux, T. Degry, N. Deutsch, D. Deville, A. Dhar, D. Dohan, S. Dowling, S. Dunning, A. Ecoffet, A. Eleti, T. Eloundou, D. Farhi, L. Fedus, N. Felix, S. P. Fishman, J. Forte, I. Fulford, L. Gao, E. Georges, C. Gibson, V. Goel, T. Gogineni, G. Goh, R. Gontijo-Lopes, J. Gordon, M. Grafstein, S. Gray, R. Greene, J. Gross, S. S. Gu, Y. Guo, C. Hallacy, J. Han, J. Harris, Y. He, M. Heaton, J. Heidecke, C. Hesse, A. Hickey, W. Hickey, P. Hoeschele, B. Houghton, K. Hsu, S. Hu, X. Hu, J. Huizinga, S. Jain, S. Jain, J. Jang, A. Jiang, R. Jiang, H. Jin, D. Jin, S. Jomoto, B. Jonn, H. Jun, T. Kaftan, Ł. Kaiser, A. Kamali, I. Kanitscheider, N. S. Keskar, T. Khan, L. Kilpatrick, J. W. Kim, C. Kim, Y. Kim, J. H. Kirchner, J. Kiros, M. Knight, D. Kokotajlo, Ł. Kondraciuk, A. Kondrich, A. Konstantinidis, K. Kosic, G. Krueger, V. Kuo, M. Lampe, I. Lan, T. Lee, J. Leike, J. Leung, D. Levy, C. M. Li, R. Lim, M. Lin, S. Lin, M. Litwin, T. Lopez, R. Lowe, P. Lue, A. Makanju, K. Malfacini, S. Manning, T. Markov, Y. Markovski, B. Martin, K. Mayer, A. Mayne, B. McGrew, S. M. McKinney, C. McLeavey, P. McMillan, J. McNeil, D. Medina, A. Mehta, J. Menick, L. Metz, A. Mishchenko, P. Mishkin, V. Monaco, E. Morikawa, D. Mossing, T. Mu, M. Murati, O. Murk, D. Mély, A. Nair, R. Nakano, R. Nayak, A. Neelakantan, R. Ngo, H. Noh, L. Ouyang, C. O'Keefe, J. Pachocki, A. Paino, J. Palermo, A. Pantuliano, G. Parascandolo, J. Parish, E. Parparita, A. Passos, M. Pavlov, A. Peng, A. Perelman, F. d. A. B. Peres, M. Petrov, H. P. d. O. Pinto, Michael, Pokorny, M. Pokrass, V. H. Pong, T. Powell, A. Power, B. Power, E. Proehl, R. Puri, A. Radford, J. Rae, A. Ramesh, C. Raymond, F. Real, K. Rimbach, C. Ross, B. Rotsted, H. Roussez, N. Ryder, M. Saltarelli, T. Sanders, S. Santurkar, G. Sastry, H. Schmidt, D. Schnurr, J. Schulman, D. Selsam, K. Sheppard, T. Sherbakov, J. Shieh, S. Shoker, P. Shyam, S. Sidor, E. Sigler, M. Simens, J. Sitkin, K. Slama, I. Sohl, B. Sokolowsky, Y. Song, N. Staudacher, F. P. Such, N. Summers, I. Sutskever, J. Tang, N. Tezak, M. B. Thompson, P. Tillet, A. Tootoonchian, E. Tseng, P. Tuggle, N. Turley, J. Tworek, J. F. C. Uribe, A. Vallone, A. Vijayvergiya, C. Voss, C. Wainwright, J. J. Wang, A. Wang, B. Wang, J. Ward, J. Wei, C. J. Weinmann, A. Welihinda, P. Welinder, J. Weng, L. Weng, M. Wiethoff, D. Willner, C. Winter, S. Wolrich, H. Wong, L. Workman, S. Wu, J. Wu, M. Wu, K. Xiao, T. Xu, S. Yoo, K. Yu, Q. Yuan, W. Zaremba, R. Zellers, C. Zhang, M. Zhang, S. Zhao, T. Zheng, J. Zhuang, W. Zhuk, and B. Zoph. GPT-4 Technical Report, Mar. 2024. URL http://arxiv.org/abs/2303.08774.

[259] Organizers of QueerInAI and A. Subramonian. NAIAC Briefing, 2023. URL https://www.queerinai.com/naiac-briefing.

[260] A. Ovalle, P. Goyal, J. Dhamala, Z. Jaggers, K.-W. Chang, A. Galstyan, R. Zemel, and R. Gupta. "I'm fully who I am": Towards Centering Transgender and Non-Binary Voices to Measure Biases in Open Language Generation. In *2023 ACM Conference on Fairness, Accountability, and Transparency*, pages 1246–1266, Chicago IL USA, June 2023. ACM. ISBN 9798400701924. doi: 10.1145/3593013.3594078. URL https://dl.acm.org/doi/10.1145/3593013.3594078.

[261] A. Ovalle, A. Subramonian, V. Gautam, G. Gee, and K.-W. Chang. Factoring the Matrix of Domination: A Critical Review and Reimagination of Intersectionality in AI Fairness. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, pages 496–511, Montr\'{e}al QC Canada, Aug. 2023. ACM. ISBN 9798400702310. doi: 10.1145/3600211.3604705. URL https://dl.acm.org/doi/10.1145/3600211.3604705.

[262] R. Parasuraman and V. Riley. Humans and Automation: Use, Misuse, Disuse, Abuse. *Human Factors*, 39(2):230–253, June 1997. ISSN 0018-7208. doi: 10.1518/001872097778543886. URL https://doi.org/10.1518/001872097778543886.

[263] P. S. Park, S. Goldstein, A. O'Gara, M. Chen, and D. Hendrycks. AI Deception: A Survey of Examples, Risks, and Potential Solutions, Aug. 2023. URL http://arxiv.org/abs/2308.14752.

[264] A. Parrish, A. Chen, N. Nangia, V. Padmakumar, J. Phang, J. Thompson, P. M. Htut, and S. Bowman. BBQ: A hand-built bias benchmark for question answering. In S. Muresan, P. Nakov, and A. Villavicencio, editors, *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2086–2105, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-acl.165. URL https://aclanthology.org/2022.findings-acl.165.

[265] J. Paschen. Investigating the emotional appeal of fake news using artificial intelligence and human contributions. *Journal of Product & Brand Management*, 29(2):223–233, May 2019. ISSN 1061-0421, 1061-0421. doi: 10.1108/JPBM-12-2018-2179. URL https://www.emerald.com/insight/content/doi/10.1108/JPBM-12-2018-2179/full/html.

[266] S. Passi and M. Vorvoreanu. Overreliance on AI: Literature Review. Technical report, Microsoft Corporation, June 2022. URL https://www.microsoft.com/en-us/research/publication/overreliance-on-ai-literature-review/.

[267] G. Pennycook, T. D. Cannon, and D. G. Rand. Prior exposure increases perceived accuracy of fake news. *Journal of Experimental Psychology: General*, 147(12):1865–1880, Dec. 2018. ISSN 1939-2222, 0096-3445. doi: 10.1037/xge0000465. URL https://doi.apa.org/doi/10.1037/xge0000465.

[268] G. Pennycook, Z. Epstein, M. Mosleh, A. A. Arechar, D. Eckles, and D. G. Rand. Shifting attention to accuracy can reduce misinformation online. *Nature*, 592(7855):590–595, Apr. 2021. ISSN 0028-0836, 1476-4687. doi: 10.1038/s41586-021-03344-2. URL https://www.nature.com/articles/s41586-021-03344-2.

[269] B. Perrigo. Inside Facebook's African Sweatshop. *Time Magazine*, 2022. URL https://time.com/6147458/facebook-africa-content-moderation-employee-treatment/.

[270] Perspective API. Perspective API. URL https://www.perspectiveapi.com/#/home.

[271] M. Phuong, M. Aitchison, E. Catt, S. Cogan, A. Kaskasoli, V. Krakovna, D. Lindner, M. Rahtz, Y. Assael, S. Hodkinson, H. Howard, T. Lieberum, R. Kumar, M. A. Raad, A. Webson, L. Ho, S. Lin, S. Farquhar, M. Hutter, G. Deletang, A. Ruoss, S. El-Sayed, S. Brown, A. Dragan, R. Shah, A. Dafoe, and T. Shevlane. Evaluating Frontier Models for Dangerous Capabilities, Apr. 2024. URL http://arxiv.org/abs/2403.13793.

[272] V. W. Polonski. AI trust and AI fears: A media debate that could divide society, 2018. URL https://www.oii.ox.ac.uk/news-events/ai-trust-and-ai-fears-a-media-debate-that-could-divide-society.

[273] B. Potts. Frontiers of multimodal learning: A responsible AI approach, Sept. 2023. URL https://www.microsoft.com/en-us/research/blog/frontiers-of-multimodal-learning-a-responsible-ai-approach/.

[274] L. Pozzobon, B. Ermis, P. Lewis, and S. Hooker. On the Challenges of Using Black-Box APIs for Toxicity Evaluation in Research, Apr. 2023. URL http://arxiv.org/abs/2304.12397.

[275] V. Prabhakaran, Z. Waseem, S. Akiwowo, and B. Vidgen. Online Abuse and Human Rights: WOAH Satellite Session at RightsCon 2020. In *Proceedings of the Fourth Workshop on Online Abuse and Harms*, pages 1–6, Online, 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.alw-1.1. URL https://www.aclweb.org/anthology/2020.alw-1.1.

[276] V. Prabhakaran, A. Mostafazadeh Davani, and M. Diaz. On Releasing Annotator-Level Labels and Information in Datasets. In *Proceedings of The Joint 15th Linguistic Annotation Workshop (LAW) and 3rd Designing Meaning Representations (DMR) Workshop*, pages 133–138, Punta Cana, Dominican Republic, 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.law-1.14. URL https://aclanthology.org/2021.law-1.14.

[277] A. Priyanshu, S. Vijay, A. Kumar, R. Naidu, and F. Mireshghallah. Are Chatbots Ready for Privacy-Sensitive Applications? An Investigation into Input Regurgitation and Prompt-Induced Sanitization, May 2023. URL http://arxiv.org/abs/2305.15008.

[278] R. Qadri, R. Shelby, C. L. Bennett, and E. Denton. AI's Regimes of Representation: A Community-centered Study of Text-to-Image Models in South Asia. In *2023 ACM Conference on Fairness, Accountability, and Transparency*, pages 506–517, Chicago IL USA, June 2023. ACM. ISBN 9798400701924. doi: 10.1145/3593013.3594016. URL https://dl.acm.org/doi/10.1145/3593013.3594016.

[279] S. Quach, P. Thaichon, K. D. Martin, S. Weaven, and R. W. Palmatier. Digital technologies: Tensions in privacy and data. *Journal of the Academy of Marketing Science*, 50(6):1299–1323, Nov. 2022. ISSN 1552-7824. doi: 10.1007/s11747-022-00845-y. URL https://doi.org/10.1007/s11747-022-00845-y.

[280] O. O. Queerinai, A. Ovalle, A. Subramonian, A. Singh, C. Voelcker, D. J. Sutherland, D. Locatelli, E. Breznik, F. Klubicka, H. Yuan, H. J, H. Zhang, J. Shriram, K. Lehman, L. Soldaini, M. Sap, M. P. Deisenroth, M. L. Pacheco, M. Ryskina, M. Mundt, M. Agarwal, N. Mclean, P. Xu, A. Pranav, R. Korpan, R. Ray, S. Mathew, S. Arora, S. John, T. Anand, V. Agrawal, W. Agnew, Y. Long, Z. J. Wang, Z. Talat, A. Ghosh, N. Dennler, M. Noseworthy, S. Jha, E. Baylor, A. Joshi, N. Y. Bilenko, A. Mcnamara, R. Gontijo-Lopes, A. Markham, E. Dong, J. Kay, M. Saraswat, N. Vytla, and L. Stark. Queer In AI: A Case Study in Community-Led Participatory AI. In *2023 ACM Conference on Fairness, Accountability, and Transparency*, pages 1882–1895, Chicago IL USA, June 2023. ACM. ISBN 9798400701924. doi: 10.1145/3593013.3594134. URL https://dl.acm.org/doi/10.1145/3593013.3594134.

[281] I. D. Raji, A. Smart, R. N. White, M. Mitchell, T. Gebru, B. Hutchinson, J. Smith-Loud, D. Theron, and P. Barnes. 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*, pages 33–44, Barcelona Spain, Jan. 2020. ACM. ISBN 978-1-4503-6936-7. doi: 10.1145/3351095.3372873. URL https://dl.acm.org/doi/10.1145/3351095.3372873.

[282] I. D. Raji, E. M. Bender, A. Paullada, E. Denton, and A. Hanna. AI and the Everything in the Whole Wide World Benchmark. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*. Curran, 2021. URL https://datasets-benchmarks-proceedings.neurips.cc/paper_files/paper/2021/file/084b6fbb10729ed4da8c3d3f5a3ae7c9-Paper-round2.pdf.

[283] A. Ramesh, M. Pavlov, G. Goh, S. Gray, C. Voss, A. Radford, M. Chen, and I. Sutskever. Zero-shot text-to-image generation. In M. Meila and T. Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8821–8831. PMLR, 2021-07-18/2021-07-24. URL https://proceedings.mlr.press/v139/ramesh21a.html.

[284] P. Rao and M. Taboada. Gender Bias in the News: A Scalable Topic Modelling and Visualization Framework. *Frontiers in Artificial Intelligence*, 4:664737, June 2021. ISSN 2624-8212. doi: 10.3389/frai.2021.664737. URL https://www.frontiersin.org/articles/10.3389/frai.2021.664737/full.

[285] A. Rapp, L. Curti, and A. Boldi. The human side of human-chatbot interaction: A systematic literature review of ten years of research on text-based chatbots. *International Journal of Human-Computer Studies*, 151:102630, July 2021. ISSN 1071-5819. doi: 10.1016/j.ijhcs.2021.102630. URL https://www.sciencedirect.com/science/article/pii/S1071581921000483.

[286] M. Rauh, J. Mellor, J. Uesato, P.-S. Huang, J. Welbl, L. Weidinger, S. Dathathri, A. Glaese, G. Irving, I. Gabriel, W. Isaac, and L. A. Hendricks. Characteristics of harmful text: Towards rigorous benchmarking of language models. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA, 2024. Curran Associates Inc. ISBN 978-1-71387-108-8.

[287] T. Ray. Common but Different Futures: AI Inequity and Climate Change, 2021. URL https://www.orfonline.org/research/common-but-different-futures-ai-inequity-and-climate-change.

[288] S. Read and I. Shine. You've probably heard of Scope 1, 2 and 3 emissions, but what are Scope 4 emissions?, Sept. 2022. URL https://www.weforum.org/agenda/2022/09/scope-4-emissions-climate-greenhouse-business/.

[289] D. Reisman, J. Schultz, K. Crawford, and M. Whittaker. Algorithmic Impact Assessments Report: A Practical Framework for Public Agency Accountability, Apr. 2018. URL https://ainowinstitute.org/publication/algorithmic-impact-assessments-report-2.

[290] N. Rekabsaz, R. West, J. Henderson, and A. Hanbury. Measuring Societal Biases from Text Corpora with Smoothed First-Order Co-occurrence. *Proceedings of the International AAAI Conference on Web and Social Media*, 15:549–560, May 2021. ISSN 2334-0770, 2162-3449. doi: 10.1609/icwsm.v15i1.18083. URL https://ojs.aaai.org/index.php/ICWSM/article/view/18083.

[291] Y. D. D.-N.-. Rep. Clarke. Algorithmic Accountability Act of 2022, Feb. 2022. URL https://www.congress.gov/bill/117th-congress/house-bill/6580/text.

[292] Republic of Korea. Input by the Government of the Republic of Korea on the Themes of an Expert Consultation on the Practical Application of the United Nations Guiding Principles on Business and Human Rights to the Activities of Technology Companies. Technical report, 2022. URL https://www.ohchr.org/sites/default/files/2022-03/RepublicofKorea.pdf.

[293] J. Ricker, S. Damm, T. Holz, and A. Fischer. Towards the Detection of Diffusion Model Deepfakes:. In *Proceedings of the 19th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications*, pages 446–457, Rome, Italy, 2024. SCITEPRESS - Science and Technology Publications. ISBN 978-989-758-679-8. doi: 10.5220/0012422000003660. URL https://www.scitepress.org/DigitalLibrary/Link.aspx?doi=10.5220/0012422000003660.

[294] R. Righi, S. Samoili, M. López Cobo, M. Vázquez-Prada Baillet, M. Cardona, and G. De Prato. The AI techno-economic complex System: Worldwide landscape, thematic subdomains and technological collaborations. *Telecommunications Policy*, 44(6):101943, July 2020. ISSN 0308-5961. doi: 10.1016/j.telpol.2020.101943. URL https://www.sciencedirect.com/science/article/pii/S0308596120300355.

[295] S. T. Roberts. *Behind the Screen: Content Moderation in the Shadows of Social Media*. Yale University Press, New Haven, 2019. ISBN 978-0-300-23588-3.

[296] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10684–10695, June 2022.

[297] J. Sablosky. "Dangerous organizations: Facebook's content moderation decisions and ethnic visibility in Myanmar". *Media, Culture & Society*, 43(6):1017–1042, Sept. 2021. ISSN 0163-4437. doi: 10.1177/0163443720987751. URL https://doi.org/10.1177/0163443720987751.

[298] V. S. Sadasivan, A. Kumar, S. Balasubramanian, W. Wang, and S. Feizi. Can AI-Generated Text be Reliably Detected?, Feb. 2024. URL http://arxiv.org/abs/2303.11156.

[299] S. Safavi, H. Gan, I. Mporas, and R. Sotudeh. Fraud Detection in Voice-Based Identity Authentication Applications and Services. In *2016 IEEE 16th International Conference on Data Mining Workshops (ICDMW)*, pages 1074–1081, Barcelona, Spain, Dec. 2016. IEEE. ISBN 978-1-5090-5910-2. doi: 10.1109/ICDMW.2016.0155. URL http://ieeexplore.ieee.org/document/7836786/.

[300] C. Saharia, W. Chan, S. Saxena, L. Li, J. Whang, E. Denton, S. K. S. Ghasemipour, B. K. Ayan, S. S. Mahdavi, R. G. Lopes, T. Salimans, J. Ho, D. J. Fleet, and M. Norouzi. Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding, May 2022. URL http://arxiv.org/abs/2205.11487.

[301] N. Sambasivan, E. Arnesen, B. Hutchinson, T. Doshi, and V. Prabhakaran. Re-imagining Algorithmic Fairness in India and Beyond. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 315–328, Virtual Event Canada, Mar. 2021. ACM. ISBN 978-1-4503-8309-7. doi: 10.1145/3442188.3445896. URL https://dl.acm.org/doi/10.1145/3442188.3445896.

[302] S. Santurkar, E. Durmus, F. Ladhak, C. Lee, P. Liang, and T. Hashimoto. Whose Opinions Do Language Models Reflect? 2023. doi: 10.48550/ARXIV.2303.17548. URL https://arxiv.org/abs/2303.17548.

[303] M. Sap, D. Card, S. Gabriel, Y. Choi, and N. A. Smith. The Risk of Racial Bias in Hate Speech Detection. In A. Korhonen, D. Traum, and L. Màrquez, editors, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1668–1678, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1163. URL https://aclanthology.org/P19-1163.

[304] M. Sap, S. Swayamdipta, L. Vianna, X. Zhou, Y. Choi, and N. A. Smith. Annotators with Attitudes: How Annotator Beliefs And Identities Bias Toxic Language Detection. In M. Carpuat, M.-C. de Marneffe, and I. V. Meza Ruiz, editors, *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5884–5906, Seattle, United States, July 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.naacl-main.431. URL https://aclanthology.org/2022.naacl-main.431.

[305] M. K. Scheuerman, A. Hanna, and E. Denton. Do Datasets Have Politics? Disciplinary Values in Computer Vision Dataset Development. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW2):1–37, Oct. 2021. ISSN 2573-0142. doi: 10.1145/3476058. URL https://dl.acm.org/doi/10.1145/3476058.

[306] K. Schoenberg. History Can Help Us Chart AI's Future, Jan. 2024. URL https://issues.org/ai-history-future-li/.

[307] P. Schramowski, M. Brack, B. Deiseroth, and K. Kersting. Safe Latent Diffusion: Mitigating Inappropriate Degeneration in Diffusion Models, Apr. 2023. URL http://arxiv.org/abs/2211.05105.

[308] R. Schwartz, A. Vassilev, K. Greene, L. Perine, A. Burt, and P. Hall. Towards a standard for identifying and managing bias in artificial intelligence. Technical Report NIST SP 1270, National Institute of Standards and Technology (U.S.), Gaithersburg, MD, Mar. 2022. URL https://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.1270.pdf.

[309] A. See, A. Pappu, R. Saxena, A. Yerukola, and C. D. Manning. Do Massively Pretrained Language Models Make Better Storytellers? In M. Bansal and A. Villavicencio, editors, *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 843–861, Hong Kong, China, Nov. 2019. Association for Computational Linguistics. doi: 10.18653/v1/K19-1079. URL https://aclanthology.org/K19-1079.

[310] S. Shafayat, E. Kim, J. Oh, and A. Oh. Multi-FAct: Assessing Multilingual LLMs' Multi-Regional Knowledge using FActScore, Mar. 2024. URL http://arxiv.org/abs/2402.18045.

[311] S. Shankar, Y. Halpern, E. Breck, J. Atwood, J. Wilson, and D. Sculley. No Classification without Representation: Assessing Geodiversity Issues in Open Data Sets for the Developing World, Nov. 2017. URL http://arxiv.org/abs/1711.08536.

[312] O. Sharir, B. Peleg, and Y. Shoham. The Cost of Training NLP Models: A Concise Overview, Apr. 2020. URL http://arxiv.org/abs/2004.08900.

[313] R. Shelby, S. Rismani, K. Henne, Aj. Moon, N. Rostamzadeh, P. Nicholas, N. Yilla, J. Gallegos, A. Smart, E. Garcia, and G. Virk. Sociotechnical Harms of Algorithmic Systems: Scoping a Taxonomy for Harm Reduction, July 2023. URL http://arxiv.org/abs/2210.05791.

[314] S. Shen, M. Zhang, W. Chen, A. Bialkowski, and M. Xu. Words Can Be Confusing: Stereotype Bias Removal in Text Classification at the Word Level. In H. Kashima, T. Ide, and W.-C. Peng, editors, *Advances in Knowledge Discovery and Data Mining*, pages 99–111, Cham, 2023. Springer Nature Switzerland. ISBN 978-3-031-33383-5. doi: 10.1007/978-3-031-33383-5_8.

[315] J. Shi, Y. Liu, P. Zhou, and L. Sun. BadGPT: Exploring Security Vulnerabilities of ChatGPT via Backdoor Attacks to InstructGPT, Feb. 2023. URL http://arxiv.org/abs/2304.12298.

[316] R. J. Shiller. *Narrative Economics: How Stories Go Viral & Drive Major Economic Events*. Princeton University Press, Princeton, 2019. ISBN 978-0-691-18229-2.

[317] K. Shilton, E. Moss, S. A. Gilbert, M. J. Bietz, C. Fiesler, J. Metcalf, J. Vitak, and M. Zimmer. Excavating awareness and power in data science: A manifesto for trustworthy pervasive data research. *Big Data & Society*, 8(2):20539517211040759, July 2021. ISSN 2053-9517. doi: 10.1177/20539517211040759. URL https://doi.org/10.1177/20539517211040759.

[318] J. Shin and S. Chan-Olmsted. User perceptions and trust of explainable machine learning fake news detectors. *International Journal of Communication*, 17(0), 2022. ISSN 1932-8036. URL https://ijoc.org/index.php/ijoc/article/view/19534.

[319] G. Simmons. Moral Mimicry: Large Language Models Produce Moral Rationalizations Tailored to Political Identity. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop)*, pages 282–297, Toronto, Canada, 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-srw.40. URL https://aclanthology.org/2023.acl-srw.40.

[320] A. Simpson. On Ethnographic Refusal: Indigeneity, 'Voice' and Colonial Citizenship. *Junctures: The Journal for Thematic Dialogue*, (9), 2007. ISSN 1179-8912. URL https://junctures.org/index.php/junctures/article/view/66.

[321] I. Solaiman. The Gradient of Generative AI Release: Methods and Considerations. In *2023 ACM Conference on Fairness, Accountability, and Transparency*, pages 111–122, Chicago IL USA, June 2023. ACM. ISBN 9798400701924. doi: 10.1145/3593013.3593981. URL https://dl.acm.org/doi/10.1145/3593013.3593981.

[322] I. Solaiman and C. Dennison. Process for Adapting Language Models to Society (PALMS) with Values-Targeted Datasets. In *Advances in Neural Information Processing Systems*, volume 34, pages 5861–5873. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper/2021/hash/2e855f9489df0712b4bd8ea9e2848c5a-Abstract.html.

[323] I. Solaiman, M. Brundage, J. Clark, A. Askell, A. Herbert-Voss, J. Wu, A. Radford, G. Krueger, J. W. Kim, S. Kreps, M. McCain, A. Newhouse, J. Blazakis, K. McGuffie, and J. Wang. Release Strategies and the Social Impacts of Language Models, Nov. 2019. URL http://arxiv.org/abs/1908.09203.

[324] D. Solove and D. K. Citron. Privacy Harms. *GW Law Faculty Publications & Other Works*, Jan. 2021. URL https://scholarship.law.gwu.edu/faculty_publications/1534.

[325] D. J. Solove. A Taxonomy of Privacy. *University of Pennsylvania Law Review*, 154(3): 477–564, 2006. ISSN 0041-9907. doi: 10.2307/40041279. URL https://www.jstor.org/stable/40041279.

[326] Spawning AI. Have I Been Trained. URL https://haveibeentrained.com/.

[327] G. Spitale, N. Biller-Andorno, and F. Germani. AI model GPT-3 (dis)informs us better than humans. *Science Advances*, 9(26):eadh1850, June 2023. ISSN 2375-2548. doi: 10.1126/sciadv.adh1850. URL https://www.science.org/doi/10.1126/sciadv.adh1850.

[328] A. Srivastava, A. Rastogi, A. Rao, A. A. M. Shoeb, A. Abid, A. Fisch, A. R. Brown, A. Santoro, A. Gupta, A. Garriga-Alonso, A. Kluska, A. Lewkowycz, A. Agarwal, A. Power, A. Ray, A. Warstadt, A. W. Kocurek, A. Safaya, A. Tazarv, A. Xiang, A. Parrish, A. Nie, A. Hussain, A. Askell, A. Dsouza, A. Slone, A. Rahane, A. S. Iyer, A. Andreassen, A. Madotto, A. Santilli, A. Stuhlmüller, A. Dai, A. La, A. Lampinen, A. Zou, A. Jiang, A. Chen, A. Vuong, A. Gupta, A. Gottardi, A. Norelli, A. Venkatesh, A. Gholamidavoodi, A. Tabassum, A. Menezes, A. Kirubarajan, A. Mullokandov, A. Sabharwal, A. Herrick, A. Efrat, A. Erdem, A. Karakaş, B. R. Roberts, B. S. Loe, B. Zoph, B. Bojanowski, B. Özyurt, B. Hedayatnia, B. Neyshabur, B. Inden, B. Stein, B. Ekmekci, B. Y. Lin, B. Howald, B. Orinion, C. Diao, C. Dour, C. Stinson, C. Argueta, C. F. Ramírez, C. Singh, C. Rathkopf, C. Meng, C. Baral, C. Wu, C. Callison-Burch, C. Waites, C. Voigt, C. D. Manning, C. Potts, C. Ramirez, C. E. Rivera, C. Siro, C. Raffel, C. Ashcraft, C. Garbacea, D. Sileo, D. Garrette, D. Hendrycks, D. Kilman, D. Roth, D. Freeman, D. Khashabi, D. Levy, D. M. González, D. Perszyk, D. Hernandez, D. Chen, D. Ippolito, D. Gilboa, D. Dohan, D. Drakard, D. Jurgens, D. Datta, D. Ganguli, D. Emelin, D. Kleyko, D. Yuret, D. Chen, D. Tam, D. Hupkes, D. Misra, D. Buzan, D. C. Mollo, D. Yang, D.-H. Lee, D. Schrader, E. Shutova, E. D. Cubuk, E. Segal, E. Hagerman, E. Barnes, E. Donoway, E. Pavlick, E. Rodola, E. Lam, E. Chu, E. Tang, E. Erdem, E. Chang, E. A. Chi, E. Dyer, E. Jerzak, E. Kim, E. E. Manyasi, E. Zheltonozhskii, F. Xia, F. Siar, F. Martínez-Plumed, F. Happé, F. Chollet, F. Rong, G. Mishra, G. I. Winata, G. de Melo, G. Kruszewski, G. Parascandolo, G. Mariani, G. Wang, G. Jaimovitch-López, G. Betz, G. Gur-Ari, H. Galijasevic, H. Kim, H. Rashkin, H. Hajishirzi, H. Mehta, H. Bogar, H. Shevlin, H. Schütze, H. Yakura, H. Zhang, H. M. Wong, I. Ng, I. Noble, J. Jumelet, J. Geissinger, J. Kernion, J. Hilton, J. Lee, J. F. Fisac, J. B. Simon, J. Koppel, J. Zheng, J. Zou, J. Kocoń, J. Thompson, J. Wingfield, J. Kaplan, J. Radom, J. Sohl-Dickstein, J. Phang, J. Wei, J. Yosinski, J. Novikova, J. Bosscher, J. Marsh, J. Kim, J. Taal, J. Engel, J. Alabi, J. Xu, J. Song, J. Tang, J. Waweru, J. Burden, J. Miller, J. U. Balis, J. Batchelder, J. Berant, J. Frohberg, J. Rozen, J. Hernandez-Orallo, J. Boudeman, J. Guerr, J. Jones, J. B. Tenenbaum, J. S. Rule, J. Chua, K. Kanclerz, K. Livescu, K. Krauth, K. Gopalakrishnan, K. Ignatyeva, K. Markert, K. D. Dhole, K. Gimpel, K. Omondi, K. Mathewson, K. Chiafullo, K. Shkaruta, K. Shridhar, K. McDonell, K. Richardson, L. Reynolds, L. Gao, L. Zhang, L. Dugan, L. Qin, L. Contreras-Ochando, L.-P. Morency, L. Moschella, L. Lam, L. Noble, L. Schmidt, L. He, L. O. Colón, L. Metz, L. K. Şenel, M. Bosma, M. Sap, M. ter Hoeve, M. Farooqi, M. Faruqui, M. Mazeika, M. Baturan, M. Marelli, M. Maru, M. J. R. Quintana, M. Tolkiehn, M. Giulianelli, M. Lewis, M. Potthast, M. L. Leavitt, M. Hagen, M. Schubert, M. O. Baitemirova, M. Arnaud, M. McElrath, M. A. Yee, M. Cohen, M. Gu, M. Ivanitskiy, M. Starritt, M. Strube, M. Swędrowski, M. Bevilacqua, M. Yasunaga, M. Kale, M. Cain, M. Xu, M. Suzgun, M. Walker, M. Tiwari, M. Bansal, M. Aminnaseri, M. Geva, M. Gheini, M. V. T, N. Peng, N. A. Chi, N. Lee, N. G.-A. Krakover, N. Cameron, N. Roberts, N. Doiron, N. Martinez, N. Nangia, N. Deckers, N. Muennighoff, N. S. Keskar, N. S. Iyer, N. Constant, N. Fiedel, N. Wen, O. Zhang, O. Agha, O. Elbaghdadi, O. Levy, O. Evans, P. A. M. Casares, P. Doshi, P. Fung, P. P. Liang, P. Vicol, P. Alipoormolabashi, P. Liao, P. Liang, P. Chang, P. Eckersley, P. M. Htut, P. Hwang, P. Miłkowski, P. Patil, P. Pezeshkpour, P. Oli, Q. Mei, Q. Lyu, Q. Chen, R. Banjade, R. E. Rudolph, R. Gabriel, R. Habacker, R. Risco, R. Millière, R. Garg, R. Barnes, R. A. Saurous, R. Arakawa, R. Raymaekers, R. Frank, R. Sikand, R. Novak, R. Sitelew, R. LeBras, R. Liu, R. Jacobs, R. Zhang, R. Salakhutdinov, R. Chi, R. Lee, R. Stovall, R. Teehan, R. Yang, S. Singh, S. M. Mohammad, S. Anand, S. Dillavou, S. Shleifer, S. Wiseman, S. Gruetter, S. R. Bowman, S. S. Schoenholz, S. Han, S. Kwatra, S. A. Rous, S. Ghazarian, S. Ghosh, S. Casey, S. Bischoff, S. Gehrmann, S. Schuster, S. Sadeghi, S. Hamdan, S. Zhou, S. Srivastava, S. Shi, S. Singh, S. Asaadi, S. S. Gu, S. Pachchigar, S. Toshniwal, S. Upadhyay, Shyamolima, Debnath, S. Shakeri, S. Thormeyer, S. Melzi, S. Reddy, S. P. Makini, S.-H. Lee, S. Torene, S. Hatwar, S. Dehaene, S. Divic, S. Ermon, S. Biderman, S. Lin, S. Prasad, S. T. Piantadosi, S. M. Shieber, S. Misherghi, S. Kiritchenko, S. Mishra, T. Linzen, T. Schuster, T. Li, T. Yu, T. Ali, T. Hashimoto, T.-L. Wu, T. Desbordes, T. Rothschild, T. Phan, T. Wang, T. Nkinyili, T. Schick, T. Kornev, T. Tunduny, T. Gerstenberg, T. Chang, T. Neeraj, T. Khot, T. Shultz, U. Shaham, V. Misra, V. Demberg, V. Nyamai, V. Raunak, V. Ramasesh, V. U. Prabhu, V. Padmakumar, V. Srikumar, W. Fedus, W. Saunders, W. Zhang, W. Vossen, X. Ren,

X. Tong, X. Zhao, X. Wu, X. Shen, Y. Yaghoobzadeh, Y. Lakretz, Y. Song, Y. Bahri, Y. Choi, Y. Yang, Y. Hao, Y. Chen, Y. Belinkov, Y. Hou, Y. Hou, Y. Bai, Z. Seid, Z. Zhao, Z. Wang, Z. J. Wang, Z. Wang, and Z. Wu. Beyond the Imitation Game: Quantifying and extrapolating the capabilities of language models, June 2023. URL http://arxiv.org/abs/2206.04615.

[329] Stanford Institute for Human-Centered Artificial Intelligence. Artificial Intelligence Index. URL https://aiindex.stanford.edu/.

[330] G. Stein, J. C. Cresswell, R. Hosseinzadeh, Y. Sui, B. L. Ross, V. Villecroze, Z. Liu, A. L. Caterini, J. E. T. Taylor, and G. Loaiza-Ganem. Exposing flaws of generative model evaluation metrics and their unfair treatment of diffusion models, Oct. 2023. URL http://arxiv.org/abs/2306.04675.

[331] E. Strubell, A. Ganesh, and A. McCallum. Energy and Policy Considerations for Deep Learning in NLP. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3645–3650, Florence, Italy, 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1355. URL https://www.aclweb.org/anthology/P19-1355.

[332] L. Struppek, D. Hintersdorf, F. Friedrich, M. Br, P. Schramowski, and K. Kersting. Exploiting Cultural Biases via Homoglyphs in Text-to-Image Synthesis. *Journal of Artificial Intelligence Research*, 78:1017–1068, Dec. 2023. ISSN 1076-9757. doi: 10.1613/jair.1.15388. URL http://www.jair.org/index.php/jair/article/view/15388.

[333] L. Struppek, D. Hintersdorf, and K. Kersting. Rickrolling the Artist: Injecting Backdoors into Text Encoders for Text-to-Image Synthesis. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4561–4573, Paris, France, Oct. 2023. IEEE. ISBN 9798350307184. doi: 10.1109/ICCV51070.2023.00423. URL https://ieeexplore.ieee.org/document/10377762/.

[334] A. Subramonian, X. Yuan, H. Daumé Iii, and S. L. Blodgett. It Takes Two to Tango: Navigating Conceptualizations of NLP Tasks and Measurements of Performance. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 3234–3279, Toronto, Canada, 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-acl.202. URL https://aclanthology.org/2023.findings-acl.202.

[335] L. Sun, Y. Huang, H. Wang, S. Wu, Q. Zhang, Y. Li, C. Gao, Y. Huang, W. Lyu, Y. Zhang, X. Li, Z. Liu, Y. Liu, Y. Wang, Z. Zhang, B. Vidgen, B. Kailkhura, C. Xiong, C. Xiao, C. Li, E. Xing, F. Huang, H. Liu, H. Ji, H. Wang, H. Zhang, H. Yao, M. Kellis, M. Zitnik, M. Jiang, M. Bansal, J. Zou, J. Pei, J. Liu, J. Gao, J. Han, J. Zhao, J. Tang, J. Wang, J. Vanschoren, J. Mitchell, K. Shu, K. Xu, K.-W. Chang, L. He, L. Huang, M. Backes, N. Z. Gong, P. S. Yu, P.-Y. Chen, Q. Gu, R. Xu, R. Ying, S. Ji, S. Jana, T. Chen, T. Liu, T. Zhou, W. Wang, X. Li, X. Zhang, X. Wang, X. Xie, X. Chen, X. Wang, Y. Liu, Y. Ye, Y. Cao, Y. Chen, and Y. Zhao. TrustLLM: Trustworthiness in Large Language Models, Mar. 2024. URL http://arxiv.org/abs/2401.05561.

[336] H. Suresh and J. Guttag. A Framework for Understanding Sources of Harm throughout the Machine Learning Life Cycle. In *Equity and Access in Algorithms, Mechanisms, and Optimization*, pages 1–9, – NY USA, Oct. 2021. ACM. ISBN 978-1-4503-8553-4. doi: 10.1145/3465416.3483305. URL https://dl.acm.org/doi/10.1145/3465416.3483305.

[337] Z. Talat and A. Lauscher. Back to the Future: On Potential Histories in NLP, Oct. 2022. URL http://arxiv.org/abs/2210.06245.

[338] Z. Talat, J. Thorne, and J. Bingel. Bridging the Gaps: Multi Task Learning for Domain Transfer of Hate Speech Detection: Multi-task Learning for Domain Transfer of Hate Speech Detection. In J. Golbeck, editor, *Online Harassment*, pages 29–55. Springer International Publishing, Cham, 2018. ISBN 978-3-319-78582-0 978-3-319-78583-7. doi: 10.1007/978-3-319-78583-7_3. URL https://link.springer.com/10.1007/978-3-319-78583-7_3.

[339] Z. Talat, S. Lulz, J. Bingel, and I. Augenstein. Disembodied Machine Learning: On the Illusion of Objectivity in NLP. Jan. 2021. URL http://arxiv.org/abs/2101.11974.

[340] Z. Talat, A. Névéol, S. Biderman, M. Clinciu, M. Dey, S. Longpre, S. Luccioni, M. Masoud, M. Mitchell, D. Radev, S. Sharma, A. Subramonian, J. Tae, S. Tan, D. Tunuguntla, and O. Van Der Wal. You reap what you sow: On the Challenges of Bias Evaluation Under Multilingual Settings. In *Proceedings of BigScience Episode #5 – Workshop on Challenges & Perspectives in Creating Large Language Models*, pages 26–41, virtual+Dublin, 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.bigscience-1.3. URL https://aclanthology.org/2022.bigscience-1.3.

[341] R. Taori and T. B. Hashimoto. Data feedback loops: Model-driven amplification of dataset biases. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23, ¡conf-loc¿, ¡city¿Honolulu¡/city¿, ¡state¿Hawaii¡/state¿, ¡country¿USA¡/country¿, ¡/conf-loc¿, 2023. JMLR.org.

[342] V. Taras, P. Steel, and B. L. Kirkman. Does Country Equate with Culture? Beyond Geography in the Search for Cultural Boundaries. *Management International Review*, 56(4):455–487, Aug. 2016. ISSN 0938-8249, 1861-8901. doi: 10.1007/s11575-016-0283-x. URL http://link.springer.com/10.1007/s11575-016-0283-x.

[343] G. Team, R. Anil, S. Borgeaud, Y. Wu, J.-B. Alayrac, J. Yu, R. Soricut, J. Schalkwyk, A. M. Dai, A. Hauth, K. Millican, D. Silver, S. Petrov, M. Johnson, I. Antonoglou, J. Schrittwieser, A. Glaese, J. Chen, E. Pitler, T. Lillicrap, A. Lazaridou, O. Firat, J. Molloy, M. Isard, P. R. Barham, T. Hennigan, B. Lee, F. Viola, M. Reynolds, Y. Xu, R. Doherty, E. Collins, C. Meyer, E. Rutherford, E. Moreira, K. Ayoub, M. Goel, G. Tucker, E. Piqueras, M. Krikun, I. Barr, N. Savinov, I. Danihelka, B. Roelofs, A. White, A. Andreassen, T. von Glehn, L. Yagati, M. Kazemi, L. Gonzalez, M. Khalman, J. Sygnowski, A. Frechette, C. Smith, L. Culp, L. Proleev, Y. Luan, X. Chen, J. Lottes, N. Schucher, F. Lebron, A. Rrustemi, N. Clay, P. Crone, T. Kocisky, J. Zhao, B. Perz, D. Yu, H. Howard, A. Bloniarz, J. W. Rae, H. Lu, L. Sifre, M. Maggioni, F. Alcober, D. Garrette, M. Barnes, S. Thakoor, J. Austin, G. Barth-Maron, W. Wong, R. Joshi, R. Chaabouni, D. Fatiha, A. Ahuja, R. Liu, Y. Li, S. Cogan, J. Chen, C. Jia, C. Gu, Q. Zhang, J. Grimstad, A. J. Hartman, M. Chadwick, G. S. Tomar, X. Garcia, E. Senter, E. Taropa, T. S. Pillai, J. Devlin, M. Laskin, D. d. L. Casas, D. Valter, C. Tao, L. Blanco, A. P. Badia, D. Reitter, M. Chen, J. Brennan, C. Rivera, S. Brin, S. Iqbal, G. Surita, J. Labanowski, A. Rao, S. Winkler, E. Parisotto, Y. Gu, K. Olszewska, Y. Zhang, R. Addanki, A. Miech, A. Louis, L. E. Shafey, D. Teplyashin, G. Brown, E. Catt, N. Attaluri, J. Balaguer, J. Xiang, P. Wang, Z. Ashwood, A. Briukhov, A. Webson, S. Ganapathy, S. Sanghavi, A. Kannan, M.-W. Chang, A. Stjerngren, J. Djolonga, Y. Sun, A. Bapna, M. Aitchison, P. Pejman, H. Michalewski, T. Yu, C. Wang, J. Love, J. Ahn, D. Bloxwich, K. Han, P. Humphreys, T. Sellam, J. Bradbury, V. Godbole, S. Samangooei, B. Damoc, A. Kaskasoli, S. M. R. Arnold, V. Vasudevan, S. Agrawal, J. Riesa, D. Lepikhin, R. Tanburn, S. Srinivasan, H. Lim, S. Hodkinson, P. Shyam, J. Ferret, S. Hand, A. Garg, T. L. Paine, J. Li, Y. Li, M. Giang, A. Neitz, Z. Abbas, S. York, M. Reid, E. Cole, A. Chowdhery, D. Das, D. Rogozińska, V. Nikolaev, P. Sprechmann, Z. Nado, L. Zilka, F. Prost, L. He, M. Monteiro, G. Mishra, C. Welty, J. Newlan, D. Jia, M. Allamanis, C. H. Hu, R. de Liedekerke, J. Gilmer, C. Saroufim, S. Rijhwani, S. Hou, D. Shrivastava, A. Baddepudi, A. Goldin, A. Ozturel, A. Cassirer, Y. Xu, D. Sohn, D. Sachan, R. K. Amplayo, C. Swanson, D. Petrova, S. Narayan, A. Guez, S. Brahma, J. Landon, M. Patel, R. Zhao, K. Villela, L. Wang, W. Jia, M. Rahtz, M. Giménez, L. Yeung, H. Lin, J. Keeling, P. Georgiev, D. Mincu, B. Wu, S. Haykal, R. Saputro, K. Vodrahalli, J. Qin, Z. Cankara, A. Sharma, N. Fernando, W. Hawkins, B. Neyshabur, S. Kim, A. Hutter, P. Agrawal, A. Castro-Ros, G. van den Driessche, T. Wang, F. Yang, S.-y. Chang, P. Komarek, R. McIlroy, M. Lučić, G. Zhang, W. Farhan, M. Sharman, P. Natsev, P. Michel, Y. Cheng, Y. Bansal, S. Qiao, K. Cao, S. Shakeri, C. Butterfield, J. Chung, P. K. Rubenstein, S. Agrawal, A. Mensch, K. Soparkar, K. Lenc, T. Chung, A. Pope, L. Maggiore, J. Kay, P. Jhakra, S. Wang, J. Maynez, M. Phuong, T. Tobin, A. Tacchetti, M. Trebacz, K. Robinson, Y. Katariya, S. Riedel, P. Bailey, K. Xiao, N. Ghelani, L. Aroyo, A. Slone, N. Houlsby, X. Xiong, Z. Yang, E. Gribovskaya, J. Adler, M. Wirth, L. Lee, M. Li, T. Kagohara, J. Pavagadhi, S. Bridgers, A. Bortsova, S. Ghemawat, Z. Ahmed, T. Liu, R. Powell, V. Bolina, M. Iinuma, P. Zablotskaia, J. Besley, D.-W. Chung, T. Dozat, R. Comanescu, X. Si, J. Greer, G. Su, M. Polacek, R. L. Kaufman, S. Tokumine, H. Hu, E. Buchatskaya, Y. Miao, M. Elhawaty, A. Siddhant, N. Tomasev, J. Xing, C. Greer, H. Miller, S. Ashraf, A. Roy, Z. Zhang, A. Ma, A. Filos, M. Besta, R. Blevins, T. Klimenko, C.-K. Yeh, S. Changpinyo, J. Mu, O. Chang, M. Pajarskas, C. Muir, V. Cohen, C. L. Lan,

K. Haridasan, A. Marathe, S. Hansen, S. Douglas, R. Samuel, M. Wang, S. Austin, C. Lan, J. Jiang, J. Chiu, J. A. Lorenzo, L. L. Sjösund, S. Cevey, Z. Gleicher, T. Avrahami, A. Boral, H. Srinivasan, V. Selo, R. May, K. Aisopos, L. Hussenot, L. B. Soares, K. Baumli, M. B. Chang, A. Recasens, B. Caine, A. Pritzel, F. Pavetic, F. Pardo, A. Gergely, J. Frye, V. Ramasesh, D. Horgan, K. Badola, N. Kassner, S. Roy, E. Dyer, V. Campos, A. Tomala, Y. Tang, D. E. Badawy, E. White, B. Mustafa, O. Lang, A. Jindal, S. Vikram, Z. Gong, S. Caelles, R. Hemsley, G. Thornton, F. Feng, W. Stokowiec, C. Zheng, P. Thacker, Ç. Ünlü, Z. Zhang, M. Saleh, J. Svensson, M. Bileschi, P. Patil, A. Anand, R. Ring, K. Tsihlas, A. Vezer, M. Selvi, T. Shevlane, M. Rodriguez, T. Kwiatkowski, S. Daruki, K. Rong, A. Dafoe, N. FitzGerald, K. Gu-Lemberg, M. Khan, L. A. Hendricks, M. Pellat, V. Feinberg, J. Cobon-Kerr, T. Sainath, M. Rauh, S. H. Hashemi, R. Ives, Y. Hasson, Y. Li, E. Noland, Y. Cao, N. Byrd, L. Hou, Q. Wang, T. Sottiaux, M. Paganini, J.-B. Lespiau, A. Moufarek, S. Hassan, K. Shivakumar, J. van Amersfoort, A. Mandhane, P. Joshi, A. Goyal, M. Tung, A. Brock, H. Sheahan, V. Misra, C. Li, N. Rakićević, M. Dehghani, F. Liu, S. Mittal, J. Oh, S. Noury, E. Sezener, F. Huot, M. Lamm, N. De Cao, C. Chen, G. Elsayed, E. Chi, M. Mahdieh, I. Tenney, N. Hua, I. Petrychenko, P. Kane, D. Scandinaro, R. Jain, J. Uesato, R. Datta, A. Sadovsky, O. Bunyan, D. Rabiej, S. Wu, J. Zhang, G. Vasudevan, E. Leurent, M. Alnahlawi, I. Georgescu, N. Wei, I. Zheng, B. Chan, P. G. Rabinovitch, P. Stanczyk, Y. Zhang, D. Steiner, S. Naskar, M. Azzam, M. Johnson, A. Paszke, C.-C. Chiu, J. S. Elias, A. Mohiuddin, F. Muhammad, J. Miao, A. Lee, N. Vieillard, S. Potluri, J. Park, E. Davoodi, J. Zhang, J. Stanway, D. Garmon, A. Karmarkar, Z. Dong, J. Lee, A. Kumar, L. Zhou, J. Evens, W. Isaac, Z. Chen, J. Jia, A. Levskaya, Z. Zhu, C. Gorgolewski, P. Grabowski, Y. Mao, A. Magni, K. Yao, J. Snaider, N. Casagrande, P. Suganthan, E. Palmer, G. Irving, E. Loper, M. Faruqui, I. Arkatkar, N. Chen, I. Shafran, M. Fink, A. Castaño, I. Giannoumis, W. Kim, M. Rybiński, A. Sreevatsa, J. Prendki, D. Soergel, A. Goedeckemeyer, W. Gierke, M. Jafari, M. Gaba, J. Wiesner, D. G. Wright, Y. Wei, H. Vashisht, Y. Kulizhskaya, J. Hoover, M. Le, L. Li, C. Iwuanyanwu, L. Liu, K. Ramirez, A. Khorlin, A. Cui, T. LIN, M. Georgiev, M. Wu, R. Aguilar, K. Pallo, A. Chakladar, A. Repina, X. Wu, T. van der Weide, P. Ponnapalli, C. Kaplan, J. Simsa, S. Li, O. Dousse, F. Yang, J. Piper, N. Ie, M. Lui, R. Pasumarthi, N. Lintz, A. Vijayakumar, L. N. Thiet, D. Andor, P. Valenzuela, C. Paduraru, D. Peng, K. Lee, S. Zhang, S. Greene, D. D. Nguyen, P. Kurylowicz, S. Velury, S. Krause, C. Hardin, L. Dixon, L. Janzer, K. Choo, Z. Feng, B. Zhang, A. Singhal, T. Latkar, M. Zhang, Q. Le, E. A. Abellan, D. Du, D. McKinnon, N. Antropova, T. Bolukbasi, O. Keller, D. Reid, D. Finchelstein, M. A. Raad, R. Crocker, P. Hawkins, R. Dadashi, C. Gaffney, S. Lall, K. Franko, E. Filonov, A. Bulanova, R. Leblond, V. Yadav, S. Chung, H. Askham, L. C. Cobo, K. Xu, F. Fischer, J. Xu, C. Sorokin, C. Alberti, C.-C. Lin, C. Evans, H. Zhou, A. Dimitriev, H. Forbes, D. Banarse, Z. Tung, J. Liu, M. Omernick, C. Bishop, C. Kumar, R. Sterneck, R. Foley, R. Jain, S. Mishra, J. Xia, T. Bos, G. Cideron, E. Amid, F. Piccinno, X. Wang, P. Banzal, P. Gurita, H. Noga, P. Shah, D. J. Mankowitz, A. Polozov, N. Kushman, V. Krakovna, S. Brown, M. Bateni, D. Duan, V. Firoiu, M. Thotakuri, T. Natan, A. Mohananey, M. Geist, S. Mudgal, S. Girgin, H. Li, J. Ye, O. Roval, R. Tojo, M. Kwong, J. Lee-Thorp, C. Yew, Q. Yuan, S. Bagri, D. Sinopalnikov, S. Ramos, J. Mellor, A. Sharma, A. Severyn, J. Lai, K. Wu, H.-T. Cheng, D. Miller, N. Sonnerat, D. Vnukov, R. Greig, J. Beattie, E. Caveness, L. Bai, J. Eisenschlos, A. Korchemniy, T. Tsai, M. Jasarevic, W. Kong, P. Dao, Z. Zheng, F. Liu, F. Yang, R. Zhu, M. Geller, T. H. Teh, J. Sanmiya, E. Gladchenko, N. Trdin, A. Sozanschi, D. Toyama, E. Rosen, S. Tavakkol, L. Xue, C. Elkind, O. Woodman, J. Carpenter, G. Papamakarios, R. Kemp, S. Kafle, T. Grunina, R. Sinha, A. Talbert, A. Goyal, D. Wu, D. Owusu-Afriyie, C. Du, C. Thornton, J. Pont-Tuset, P. Narayana, J. Li, S. Fatehi, J. Wieting, O. Ajmeri, B. Uria, T. Zhu, Y. Ko, L. Knight, A. Héliou, N. Niu, S. Gu, C. Pang, D. Tran, Y. Li, N. Levine, A. Stolovich, N. Kalb, R. Santamaria-Fernandez, S. Goenka, W. Yustalim, R. Strudel, A. Elqursh, B. Lakshminarayanan, C. Deck, S. Upadhyay, H. Lee, M. Dusenberry, Z. Li, X. Wang, K. Levin, R. Hoffmann, D. Holtmann-Rice, O. Bachem, S. Yue, S. Arora, E. Malmi, D. Mirylenka, Q. Tan, C. Koh, S. H. Yeganeh, S. Põder, S. Zheng, F. Pongetti, M. Tariq, Y. Sun, L. Ionita, M. Seyedhosseini, P. Tafti, R. Kotikalapudi, Z. Liu, A. Gulati, J. Liu, X. Ye, B. Chrzaszcz, L. Wang, N. Sethi, T. Li, B. Brown, S. Singh, W. Fan, A. Parisi, J. Stanton, C. Kuang, V. Koverkathu, C. A. Choquette-Choo, Y. Li, T. J. Lu, A. Ittycheriah, P. Shroff, P. Sun, M. Varadarajan, S. Bahargam, R. Willoughby, D. Gaddy, I. Dasgupta, G. Desjardins, M. Cornero, B. Robenek, B. Mittal, B. Albrecht, A. Shenoy, F. Moiseev, H. Jacobsson, A. Ghaffarkhah, M. Rivière, A. Walton, C. Crepy, A. Parrish, Y. Liu, Z. Zhou, C. Farabet, C. Radebaugh, P. Srinivasan, C. van der Salm, A. Fidjeland, S. Scellato, E. Latorre-

Chimoto, H. Klimczak-Plucińska, D. Bridson, D. de Cesare, T. Hudson, P. Mendolicchio, L. Walker, A. Morris, I. Penchev, M. Mauger, A. Guseynov, A. Reid, S. Odoom, L. Loher, V. Cotruta, M. Yenugula, D. Grewe, A. Petrushkina, T. Duerig, A. Sanchez, S. Yadlowsky, A. Shen, A. Globerson, A. Kurzrok, L. Webb, S. Dua, D. Li, P. Lahoti, S. Bhupatiraju, D. Hurt, H. Qureshi, A. Agarwal, T. Shani, M. Eyal, A. Khare, S. R. Belle, L. Wang, C. Tekur, M. S. Kale, J. Wei, R. Sang, B. Saeta, T. Liechty, Y. Sun, Y. Zhao, S. Lee, P. Nayak, D. Fritz, M. R. Vuyyuru, J. Aslanides, N. Vyas, M. Wicke, X. Ma, T. Bilal, E. Eltyshev, D. Balle, N. Martin, H. Cate, J. Manyika, K. Amiri, Y. Kim, X. Xiong, K. Kang, F. Luisier, N. Tripuraneni, D. Madras, M. Guo, A. Waters, O. Wang, J. Ainslie, J. Baldridge, H. Zhang, G. Pruthi, J. Bauer, F. Yang, R. Mansour, J. Gelman, Y. Xu, G. Polovets, J. Liu, H. Cai, W. Chen, X. Sheng, E. Xue, S. Ozair, A. Yu, C. Angermueller, X. Li, W. Wang, J. Wiesinger, E. Koukoumidis, Y. Tian, A. Iyer, M. Gurumurthy, M. Goldenson, P. Shah, M. K. Blake, H. Yu, A. Urbanowicz, J. Palomaki, C. Fernando, K. Brooks, K. Durden, H. Mehta, N. Momchev, E. Rahimtoroghi, M. Georgaki, A. Raul, S. Ruder, M. Redshaw, J. Lee, K. Jalan, D. Li, G. Perng, B. Hechtman, P. Schuh, M. Nasr, M. Chen, K. Milan, V. Mikulik, T. Strohman, J. Franco, T. Green, D. Hassabis, K. Kavukcuoglu, J. Dean, and O. Vinyals. Gemini: A Family of Highly Capable Multimodal Models, Dec. 2023. URL http://arxiv.org/abs/2312.11805.

[344] The GDELT Project. The Unintended Consequences & Harms Of Multimodal LLM Debiasing: Detection Vs Generation, 2023. URL https://blog.gdeltproject.org/the-unintended-consequences-harms-of-multimodal-llm-debiasing-detection-vs-generation/.

[345] The Glaze Project, S. Shan, B. Zhao, H. Zheng, W. Ding, A. Ha, J. Passananti, S. Wu, R. Bhaskar, J. Cryan, and E. Wenger. Glaze - Protecting Artists from Generative AI. URL https://glaze.cs.uchicago.edu/.

[346] The International Association of Privacy Professionals. Global Privacy Law and DPA Directory, 2024. URL https://iapp.org/resources/global-privacy-directory/.

[347] The Ministry of Economy, Trade and Industry. Governance Guidelines for Implementation of AI Principles Ver. 1.1" Compiled, 2022. URL https://www.meti.go.jp/press/2021/01/20220125001/20220124003.html.

[348] D. Thiel, M. Stroebel, and R. Portnoff. Generative ML and CSAM: Implications and Mitigations. 2023. doi: 10.25740/jv206yg3793. URL https://purl.stanford.edu/jv206yg3793.

[349] Thorn. Generative AI: Now is the Time for Safety By Design, May 2023. URL https://www.thorn.org/blog/now-is-the-time-for-safety-by-design/.

[350] N. Thylstrup and Z. Talat. Detecting 'Dirt' and 'Toxicity': Rethinking Content Moderation as Pollution Behaviour. *SSRN Electronic Journal*, 2020. ISSN 1556-5068. doi: 10.2139/ssrn.3709719. URL https://www.ssrn.com/abstract=3709719.

[351] N. Todorovic and A. Chaudhuri. Using AI to help organizations detect and report child sexual abuse material online, Sept. 2018. URL https://blog.google/around-the-globe/google-europe/using-ai-help-organizations-detect-and-report-child-sexual-abuse-material-online/.

[352] J. A. Tomain. Online Privacy and the First Amendment: An Opt-In Approach to Data Processing, Feb. 2014. URL https://papers.ssrn.com/abstract=2573206.

[353] J. Tomlinson. Cultural Imperialism. In G. Ritzer, editor, *The Wiley-Blackwell Encyclopedia of Globalization*. Wiley, 1 edition, Feb. 2012. ISBN 978-1-4051-8824-1 978-0-470-67059-0. doi: 10.1002/9780470670590.wbeog129. URL https://onlinelibrary.wiley.com/doi/10.1002/9780470670590.wbeog129.

[354] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, A. Rodriguez, A. Joulin, E. Grave, and G. Lample. LLaMA: Open and Efficient Foundation Language Models, Feb. 2023. URL http://arxiv.org/abs/2302.13971.

[355] Treasury Board of Canada Secretariat. Algorithmic Impact Assessment Tool, Mar. 2021. URL https://www.canada.ca/en/government/system/digital-government/digital-government-innovations/responsible-use-ai/algorithmic-impact-assessment.html.

[356] UN Human Rights Office of the High Commissioner. Non-discrimination (ND) Enhancing equality and countering discrimination, 2021. URL https://www2.ohchr.org/english/ohchrreport2021/documents/Non-DiscriminationND.pdf.

[357] United Kingdom's Department for Digital, Culture, Media & Sport and D. Collins. UK sets out proposals for new AI rulebook to unleash innovation and boost public trust in the technology, 2022. URL https://www.gov.uk/government/news/uk-sets-out-proposals-for-new-ai-rulebook-to-unleash-innovation-and-boost-public-trust-in-the-technology.

[358] United Nations Framework Convention on Climate Change. Methodologies and Tools to Evaluate Climate Change Impacts and Adaptation. URL https://unfccc.int/methodologies-and-tools-to-evaluate-climate-change-impacts-and-adaptation-2.

[359] U.S. Army Acquisition Support Center. U.S. army acquisition support center, 2024. URL https://asc.army.mil/web/news-generation-generation/.

[360] U.S. Department of State. Political Declaration on Responsible Military Use of Artificial Intelligence and Autonomy, 2024. URL https://www.state.gov/political-declaration-on-responsible-military-use-of-artificial-intelligence-and-autonomy/.

[361] C. Vaccari and A. Chadwick. Deepfakes and Disinformation: Exploring the Impact of Synthetic Political Video on Deception, Uncertainty, and Trust in News. *Social Media + Society*, 6(1): 205630512090340, Jan. 2020. ISSN 2056-3051, 2056-3051. doi: 10.1177/2056305120903408. URL http://journals.sagepub.com/doi/10.1177/2056305120903408.

[362] M. Veale and R. Binns. Fairer machine learning in the real world: Mitigating discrimination without collecting sensitive data. *Big Data & Society*, 4(2):205395171774353, Dec. 2017. ISSN 2053-9517, 2053-9517. doi: 10.1177/2053951717743530. URL http://journals.sagepub.com/doi/10.1177/2053951717743530.

[363] A. Venigalla and L. Li. Mosaic LLMs: GPT-3 quality for, Thu, 09/29/2022 - 13:28. URL https://www.databricks.com/blog/gpt-3-quality-for-500k.

[364] B. Vidgen, A. Agrawal, A. M. Ahmed, V. Akinwande, N. Al-Nuaimi, N. Alfaraj, E. Alhajjar, L. Aroyo, T. Bavalatti, B. Blili-Hamelin, et al. Introducing v0. 5 of the ai safety benchmark from mlcommons. *arXiv preprint arXiv:2404.12241*, 2024.

[365] J. Vipra and S. M. West. Computational Power and AI. Technical report, AI Now Institute, New York, NY, USA, 2023. URL https://ainowinstitute.org/publication/policy/compute-and-ai.

[366] D. Walsh. The legal issues presented by generative AI, 2023. URL https://mitsloan.mit.edu/ideas-made-to-matter/legal-issues-presented-generative-ai.

[367] A. Wang, A. Singh, J. Michael, F. Hill, O. Levy, and S. Bowman. GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium, 2018. Association for Computational Linguistics. doi: 10.18653/v1/W18-5446. URL http://aclweb.org/anthology/W18-5446.

[368] A. Wang, V. V. Ramaswamy, and O. Russakovsky. Towards Intersectionality in Machine Learning: Including More Identities, Handling Underrepresentation, and Performing Evaluation. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 336–349, Seoul Republic of Korea, June 2022. ACM. ISBN 978-1-4503-9352-2. doi: 10.1145/3531146.3533101. URL https://dl.acm.org/doi/10.1145/3531146.3533101.

[369] A. Wang, X. Bai, S. Barocas, and S. L. Blodgett. Measuring machine learning harms from stereotypes: Requires understanding who is being harmed by which errors in what ways, Feb. 2024. URL http://arxiv.org/abs/2402.04420.

[370] Z. Waseem. Are You a Racist or Am I Seeing Things? Annotator Influence on Hate Speech Detection on Twitter. In *Proceedings of the First Workshop on NLP and Computational Social Science*, pages 138–142, Austin, Texas, 2016. Association for Computational Linguistics. doi: 10.18653/v1/W16-5618. URL http://aclweb.org/anthology/W16-5618.

[371] L. Weidinger, J. Mellor, M. Rauh, C. Griffin, J. Uesato, P.-S. Huang, M. Cheng, M. Glaese, B. Balle, A. Kasirzadeh, Z. Kenton, S. Brown, W. Hawkins, T. Stepleton, C. Biles, A. Birhane, J. Haas, L. Rimell, L. A. Hendricks, W. Isaac, S. Legassick, G. Irving, and I. Gabriel. Ethical and social risks of harm from Language Models, Dec. 2021. URL http://arxiv.org/abs/2112.04359.

[372] L. Weidinger, K. R. McKee, R. Everett, S. Huang, T. O. Zhu, M. J. Chadwick, C. Summerfield, and I. Gabriel. Using the Veil of Ignorance to align AI systems with principles of justice. *Proceedings of the National Academy of Sciences*, 120(18):e2213709120, May 2023. doi: 10.1073/pnas.2213709120. URL https://www.pnas.org/doi/10.1073/pnas.2213709120.

[373] L. Weidinger, M. Rauh, N. Marchal, A. Manzini, L. A. Hendricks, J. Mateos-Garcia, S. Bergman, J. Kay, C. Griffin, B. Bariach, I. Gabriel, V. Rieser, and W. Isaac. Sociotechnical Safety Evaluation of Generative AI Systems, Oct. 2023. URL http://arxiv.org/abs/2310.11986.

[374] R. Weitzer. Racial discrimination in the criminal justice system: Findings and problems in the literature. *Journal of Criminal Justice*, 24(4):309–322, Jan. 1996. ISSN 0047-2352. doi: 10.1016/0047-2352(96)00015-3. URL https://www.sciencedirect.com/science/article/pii/0047235296000153.

[375] Y. Wen, Y. Liu, C. Chen, and L. Lyu. Detecting, Explaining, and Mitigating Memorization in Diffusion Models. In *The Twelfth International Conference on Learning Representations*, Oct. 2023. URL https://openreview.net/forum?id=84n3UwkH7b.

[376] F. Westin and S. Chiasson. Opt out of privacy or "go home": Understanding reluctant privacy behaviours through the FoMO-centric design paradigm. In *Proceedings of the New Security Paradigms Workshop*, NSPW '19, pages 57–67, New York, NY, USA, Jan. 2020. Association for Computing Machinery. ISBN 978-1-4503-7647-1. doi: 10.1145/3368860.3368865. URL https://doi.org/10.1145/3368860.3368865.

[377] C. C. Williams and A. Efendic. Evaluating the relationship between marginalization and participation in undeclared work: Lessons from Bosnia and Herzegovina. *Southeast European and Black Sea Studies*, 21(3):481–499, July 2021. ISSN 1468-3857. doi: 10.1080/14683857.2021.1928419. URL https://doi.org/10.1080/14683857.2021.1928419.

[378] L. Winner. Do Artifacts Have Politics? *Daedalus*, 109(1), 1980. URL http://www.jstor.org/stable/20024652.

[379] A. Wolfers. "National Security" as an Ambiguous Symbol. *Political Science Quarterly*, 67(4):481–502, Dec. 1952. ISSN 0032-3195, 1538-165X. doi: 10.2307/2145138. URL https://academic.oup.com/psq/article/67/4/481/7245057.

[380] C.-J. Wu, R. Raghavendra, U. Gupta, B. Acun, N. Ardalani, K. Maeng, G. Chang, F. Aga, J. Huang, C. Bai, M. Gschwind, A. Gupta, M. Ott, A. Melnikov, S. Candido, D. Brooks, G. Chauhan, B. Lee, H.-H. Lee, B. Akyildiz, M. Balandat, J. Spisak, R. Jain, M. Rabbat, and K. Hazelwood. Sustainable AI: Environmental implications, challenges and opportunities. In D. Marculescu, Y. Chi, and C. Wu, editors, *Proceedings of Machine Learning and Systems*, volume 4, pages 795–813, 2022. URL https://proceedings.mlsys.org/paper_files/paper/2022/file/462211f67c7d858f663355eff93b745e-Paper.pdf.

[381] S. Wyllie, I. Shumailov, and N. Papernot. Fairness Feedback Loops: Training on Synthetic Data Amplifies Bias, Mar. 2024. URL http://arxiv.org/abs/2403.07857.

[382] A. Xu, E. Pathak, E. Wallace, S. Gururangan, M. Sap, and D. Klein. Detoxifying Language Models Risks Marginalizing Minority Voices. In K. Toutanova, A. Rumshisky, L. Zettlemoyer, D. Hakkani-Tur, I. Beltagy, S. Bethard, R. Cotterell, T. Chakraborty, and Y. Zhou, editors, *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2390–2397, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.190. URL https://aclanthology.org/2021.naacl-main.190.

[383] Z. Yan, Y. Zhang, X. Yuan, S. Lyu, and B. Wu. DeepfakeBench: A comprehensive benchmark of deepfake detection. In A. Oh, T. Neumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 4534–4565. Curran Associates, Inc., 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/0e735e4b4f07de483cbe250130992726-Paper-Datasets_and_Benchmarks.pdf.

[384] A. G. Yasar, A. Chong, E. Dong, T. Gilbert, S. Hladikova, C. Mougan, X. Shen, S. Singh, A.-A. Stoica, and S. Thais. Integration of generative ai in the digital markets act: Contestability and fairness from a cross-disciplinary perspective. *SSRN Electronic Journal*, 2024. ISSN 1556-5068. doi: 10.2139/ssrn.4769439. URL http://dx.doi.org/10.2139/ssrn.4769439.

[385] J. Zaller and S. Feldman. A Simple Theory of the Survey Response: Answering Questions versus Revealing Preferences. *American Journal of Political Science*, 36(3):579, Aug. 1992. ISSN 00925853. doi: 10.2307/2111583. URL https://www.jstor.org/stable/2111583?origin=crossref.

[386] S. Zhang, S. Roller, N. Goyal, M. Artetxe, M. Chen, S. Chen, C. Dewan, M. Diab, X. Li, X. V. Lin, T. Mihaylov, M. Ott, S. Shleifer, K. Shuster, D. Simig, P. S. Koura, A. Sridhar, T. Wang, and L. Zettlemoyer. OPT: Open Pre-trained Transformer Language Models, June 2022. URL http://arxiv.org/abs/2205.01068.

[387] D. Zhuang, X. Zhang, S. Song, and S. Hooker. Randomness in Neural Network Training: Characterizing the Impact of Tooling. *Proceedings of Machine Learning and Systems*, 4:316–336, Apr. 2022. URL https://proceedings.mlsys.org/paper_files/paper/2022/hash/427e0e886ebf87538afdf0badb805b7f-Abstract.html.