Representation in AI Evaluations

A. Stevie Bergman*
steviebergman@deepmind.com
DeepMind
New York, USA

Lisa Anne Hendricks
lmh@deepmind.com
DeepMind
London, UK

Maribeth Rauh
mbrauh@deepmind.com
DeepMind
London, UK

Boxi Wu
boxi@deepmind.com
DeepMind
London, UK

William Agnew
wagnew3@cs.washington.edu
University of Washington
Seattle, Washington, USA

Markus Kunesch
mkunesch@deepmind.com
DeepMind
London, UK

Isabella Duan
isabelladuan@uchicago.edu
University of Chicago
Chicago, Illinois, USA

Iason Gabriel
iason@deepmind.com
DeepMind
London, UK

William Isaac
williamis@deepmind.com
DeepMind
London, UK

ABSTRACT

Calls for representation in artificial intelligence (AI) and machine learning (ML) are widespread, with "representation" or "representativeness" generally understood to be both an instrumentally and intrinsically beneficial quality of an AI system, and central to fairness concerns. But what does it mean for an AI system to be "representative"? Each element of the AI lifecycle is geared towards its own goals and effect on the system, therefore requiring its own analyses with regard to what kind of representation is best. In this work we untangle the benefits of representation in AI evaluations to develop a framework to guide an AI practitioner or auditor towards the creation of representative ML evaluations. Representation, however, is not a panacea. We further lay out the limitations and tensions of instrumentally representative datasets, such as the necessity of data existence and access, surveillance vs expectations of privacy, implications for foundation models and power. This work sets the stage for a research agenda on representation in AI, which extends beyond instrumentally valuable representation in evaluations towards refocusing on, and empowering, impacted communities.

CCS CONCEPTS

• **Computing methodologies** → **Machine learning**; **Artificial intelligence**.

KEYWORDS

datasets, responsible AI, machine learning evaluation

ACM Reference Format:
A. Stevie Bergman, Lisa Anne Hendricks, Maribeth Rauh, Boxi Wu, William Agnew, Markus Kunesch, Isabella Duan, Iason Gabriel, and William Isaac. 2023. Representation in AI Evaluations. In *2023 ACM Conference on Fairness, Accountability, and Transparency (FAccT '23), June 12–15, 2023, Chicago, IL, USA.* ACM, New York, NY, USA, 15 pages. https://doi.org/10.1145/3593013.3594019

FAccT '23, June 12–15, 2023, Chicago, IL, USA

ACM ISBN 979-8-4007-0192-4/23/06.
https://doi.org/10.1145/3593013.3594019

1 INTRODUCTION

What does it mean for an artificial intelligence (AI) system to be "representative"? In discourse on responsible machine learning (ML), "representativeness" is generally understood to be both an intrinsic and instrumental good [67], however, as with other capacious ideas in the AI ethics realm (*e.g.* AI fairness and transparency), the intuitive sense of the concept belies the fact that the term has a variety of different meanings in different contexts [26, 125]. The *Oxford Reference* has 11 definitions for representation starting with "Depicting or 'making present' something which is absent (*e.g.* people, places, events, or abstractions) in a different form" [119]. This is exemplified by the manner in which people's preferences could be captured, represented, or modeled in ML (the different form). Representation is distinct from replication, as the form or medium changes – and the "what" and "how (in what ways)" something is depicted or described will vary for different types of representations [119].[1] Even without a singular, clear and nuanced definition and best practice for machine learning, calls for representation in AI are widespread (*e.g.* [3, 31, 37, 46, 50, 80, 98, 103, 104, 134]), and representation – or lack thereof – is often held up as central to fairness in AI, if not the cause of the fairness concerns (*e.g.* representation bias and harms [118]). "Representation" is further offered in the AI fairness field as a mitigation for bias, implying that greater inclusion of underrepresented groups (*e.g.* in datasets) is an important goal. However, an oversimplified interpretation of this notion can lead to objectification and exploitation (*e.g.* [52]), rendering it all the more urgent to clarify the term's meaning, and develop a framework for understanding what constitutes good practice towards instrumentally and intrinsically valuable representation [23, 34, 37, 41, 91, 105].

In this endeavor, we must contend with the fact that there are a number of potential sites of representation or exclusion in the

[1] Formal definitions of "representation," "representative," or the verb "to represent" further include "to present again or anew," "to present by means of something standing in the place of," "to typify," "to portray visually," or "to stand in the act of" [147].

AI lifecycle [98] including through the datasets that are used to train, evaluate, audit, and/or benchmark models [25, 28, 70, 105, 122, 123, 125, 132], stakeholders who evaluate the technology [47], annotators who label datasets or probe models [15, 103], or even a property of the ML designers themselves [47, 104]. Each of these sites serves a different purpose, thus each requires its own analyses with regard to the precise formulation of representation that is truly beneficial.

Here, we focus on AI benchmark, audit, or evaluation datasets,[2] as greater representation in model evaluations has the clear instrumental value of generating more robust tests and thereby surfacing more blindspots leading to higher performing systems, and avoiding harms from burdening already marginalised communities.

While representation in training datasets has been shown to be a significant challenge [122], perhaps a still greater challenge – and one that is difficult to detect – is posed by poor representation in evaluation datasets. This can lead to insidious situations such as being unable to surface low model performance or obtaining artificially high metrics. Thus, the aim of this work is to build a practical understanding of, and stronger consensus around, representation in machine learning model evaluations. Section 2 provides a general overview of the manners in which the term "representation" has been employed in machine learning and sociotechnical fields – building an intuitive understanding of the concepts, tensions, and motivations at play. In Section 3 we build a framework for the formulation of representation in machine learning evaluations that can detect and achieve performant systems. This is meant to inform and guide practitioners, auditors, AI ethicists, and policymakers (*e.g.* [46, 98]) seeking to audit and build better ML systems. To achieve this, we combine a list of key questions to consider with examples from the literature and a review of the NeurIPS 2022 Datasets and Benchmarks track. We then bring together pieces of good practice to describe subject-domain representation. In Section 4 we lay out the limitations and tensions of the instrumental-only approach for representation taken in this paper, helping to guide thinking on what intrinsically valuable, inclusive representation could mean. We then proceed to describe a research agenda (Section 5) to build the field of necessary, structured inquiry in order to advocate for and develop effective and beneficially representative machine learning systems.

2 USES OF REPRESENTATION IN MACHINE LEARNING

As mentioned, calls for representation in AI systems are commonplace, with uses of the term often proving somewhat unclear [26], leading to dangerous cascades of misunderstanding [125]. In sociotechnical fields, "representation" is sometimes employed as a catch-all term for responsible implementation, or as a synonym for "real-world" testing and applicability [26]. At other times, "representativeness" is not mentioned at all (or is not the primary focus), but is addressed indirectly via related terms such as "diversity" or "fairness" [47]. To our knowledge the only direct attempt to confront the many definitions and implications of representation for ML is undertaken by Chasalow and Levy [26]. In their work, the authors take a primarily historic view, and surface the fact that "representativeness" is a slippery concept and a "suitcase word," due to the fact that many meanings are packed into the term [26, 93].

It is instructive to begin unpacking representation by touching upon a few pieces of fundamental scholarship. These works are primarily on the topic of political representation, with classic efforts including Pitkin's "four views of representation" [107] and Pettit's *Varieties of Public Representation*. Pettit [106] describes representation as having three parts: the *representor* (body doing the representing), *representee* (body being represented), and the connection between them, highlighting the necessarily relational nature of representation. Indeed, representation is perhaps most commonly discussed in reference to effective democratic governance. One can be represented, such as having one's vote counted or have a representative in an elected congress or parliament. In fact, a legislative body will be set up differently depending on whether it is designed to represent the interests or the identities of constituents (*e.g.* the US Senate vs polling to capture the overall interests or preferences of the US population) [152]. Similarly, when designing machine learning systems to be representative, it is incumbent upon practitioners to justify why one kind of representation is more important than another, given the wider goals and context of the system.

For a visual example of a representation, consider a photograph. The image in a photo is a two-dimensional, lower-resolution representation of our effectively infinite-resolution reality. It is also a dataset of pixels that can be either a good or bad representation of reality, depending on the goals of the photo. Following this, image compression offers an insight: representation defined by what information can be discarded. In compressing a photograph – say, of deep space – an algorithm may decide to combine the pixels between stars, leaving only the galaxies in high resolution. The image now takes up less disk space, but retains the desired information on the galaxies. However, if the primary interest is in the varying shades of deep space – which some scientists are indeed focused on – a compression that retains only the resolution *between* the stars and galaxies is needed. In this way, what is represented is what is valued.

In the mathematically-focused machine learning literature, one encounters related topics such as: representation theory, sampling of larger populations to match the distribution [55, 100], and representation learning and embeddings. In mathematics, representation theory is the study of abstract mathematical structures by exactly mimicking their behavior with matrices, whose properties are well-understood. Representation learning refers to a model learning the patterns latent in its training data, thereby "representing" that data [13]. These patterns – or representations – are expected to be shared, underlying concepts, and can form the basis for predictive models. "Representative sampling" in statistics typically refers to normal, random, non-stratified sampling where the goal is to match the distribution of a larger dataset [100]. Yet statistics and machine learning are tools for different tasks, *i.e.* population inferences vs generalisable predictive patterns [24]. Each of these topics share a practical connection to representation in evaluation datasets for AI: they aim for lower complexity or dimensionality than some larger structure, while still capturing desired behavior, *a.k.a.* depicting something which is absent in a different form [119].

[2]The delineation between evaluation, audit, and benchmark datasets is unclear. In this paper we primarily use "evaluation" datasets as an umbrella term to refer to all three.

2.1 Representation in sociotechnical literature

The concepts of *representational harms* and *representational biases* are prominent in the algorithmic fairness discourse [8, 133, 134]. Representational harms occur "when the development sample under-represents some part of the population, and subsequently fails to generalise well for a subset of the use population" [134]. Representational harms are the primary finding in several significant works in the sociotechnical field, including research surfacing low performance and under-representation of darker skinned subjects in prominent facial analysis datasets (by employing an auditing dataset that is more representative) [23], case studies demonstrating that images in large object recognition datasets have been overwhelmingly sourced from wealthier Western nations (and thus are ineffective at recognising lower socioeconomic or non-Western items and events) [37, 132], and that millions of images of people in ImageNet [40] were labeled with offensive categories including racial slurs ("problematic representations") [34]. These case studies (and others, *e.g.* [36, 72, 105, 134]) display how representation, or lack thereof, directly corresponds to poor performance. This, in turn, demonstrates the instrumental value of robust representation for AI systems.

Representational biases arise when a system (*e.g.* a search engine or a generative AI) represents some social groups in a less favorable light than others, demeans them (*e.g.* [59, 87]), or fails to recognise their existence altogether [2, 18, 19, 33, 72, 133, 143]. Representational harms are typically considered a model outcome, commonly triggered by (1) poorly representative training datasets, and/or (2) insufficiently representative testing, such that issues were not surfaced and fixed in the model to prevent representational harms to users prior to deployment [26]. Notably, poor representation in evaluation and/or training datasets can lead to representation, allocation (inequitable distribution of resources or opportunities), and capability harms (upstream performance disparity) [118]. In the case of evaluations, this could stem from a lack of visibility into whether the model is performant for marginalised populations, leading to blindspots and the possibility of cascading harms, such as those described in Elish [44].

A central argument of Chasalow and Levy [26] is that "representativeness can rarely stand alone," meaning that it will need some description or augmentation (*i.e.* an adjective) to have definition. Without it, effective communication is hampered on a concept that is vital to the proper functioning of AI systems. There is extensive literature on the types and resulting harms of poor representation [133, 134], but the question of what good representation means theoretically and practically in the context of evaluating AI models is less explored.

3 REPRESENTATION IN EVALUATION DATASETS

What is meant by a "representative evaluation"? The operational details can be subtle, with "different meanings hav[ing] distinct practical and normative implications" [26]. Any claims of representation immediately raise questions along the lines of "what" and "how (in what ways)" from the *Oxford Reference* [119]: (i) representative of what (*i.e.* what attribute)?, and (ii) of who (*i.e.* which people)? To illustrate the importance of these questions, we present two hypothetical examples below.

Example 1: Street Signs. A model is developed to detect and interpret the meaning of street signs, *e.g.* in order to help pilot autonomous vehicles. This model is meant to be released in the US, Indonesia, Japan, and the UK. It needs to perform well in all four countries before it is released.

In this example, if the evaluation dataset is entirely composed of images of US traffic signs, then the practitioner will not have visibility into whether the model will work well for Indonesia, Japan, or the UK. One *might* expect the model to work well for Indonesia and Japan as their street signs follow the MUTCD standardization [146] in the US. However, the practitioner cannot truly show high performance for Japan and Indonesia with the evaluation dataset described, and there is no reason to believe the model will work well in the UK, which does not follow MUTCD standardization.

For the simple use-case described above, key questions include: *Representative of what?* This would be the array of types of street signs and their meanings. *Representative of which groups?* Street sign types in the US, Japan, Indonesia, and the UK. With this information, we argue the necessary question to ask is: *What is robust representation for this model evaluation i.e. what kind of representation is needed?* For this example, the practitioner might consider the dataset to be representative of US street signs if it contains all the different types of street signs to be detected in the US. Similarly so for Japan, Indonesia, and the UK. Additionally, the practitioner may want to ensure the set captures attributes such as urban vs rural settings, new vs old signs, lighting, backgrounds, and other variables shown to be important to image recognition tasks.

Example 2: Portuguese Hate Speech. A model is developed to automatically detect hate speech in Portuguese, with the intention to deploy it globally. It needs to perform well everywhere Portuguese is spoken.

For this example, we return to the questions posed above, namely: *Representative of what?* In this case, the answer is uses of hate speech in Portuguese. However, the legal and practical identification of hate speech is a notoriously difficult topic, with such language often being highly contextual and difficult to disambiguate from the merely offensive [42]. *Representative of which groups?* Portuguese is the official language in ten nations, with the most populous being Brazil, Angola, Mozambique, and Portugal. Yet, there's a significant number of speakers outside of these countries, including Venezuela, Guyana, and the US. If this model is meant to detect hate speech *globally*, best practice would consider communities of Portuguese speakers in all these nations.

Last, *What is robust representation for this model evaluation?* In this application, the evaluation would need to capture the different manners in which Portuguese-speaking communities use hate speech *i.e.* the types of samples – or events – that could pass through the model in operation. Notably, the amount of content needed to represent the manners in which hate speech is used by Portuguese-speaking communities does not necessarily track their population sizes. Instead, it is important to consider the different communities who speak Portuguese, and whether there are relevant differences

between them that bear upon this model's purpose. For example, if the Portuguese spoken in Brazil and Portugal differ significantly in terms of how these communities employ hate speech,[3] then a dataset that is representative of Brazilian-Portuguese hate speech will not be representative of hate speech in the Portugal-Portuguese context.

With these examples in mind, we proceed to describe standard practice generally, then build towards *subject-domain representation* via beneficial practices that focus on sampling for data subject groups and are informed by domain.

3.1 Standard practice

Initial reflections on representative sampling often draw upon the statistical notion described in Section 2, where "representation" implies that the composition of the dataset matches the corresponding population along some measurable features or variables. For example, a language dataset might be considered "representative" if the distribution of languages in the set is proportional to the population of communities whose language is being modeled. As discussed, uniform – or random – sampling is useful for inference (*e.g.* political polling) however it is not necessarily the right practice for effective AI evaluation [24].

Standard practice in the creation of datasets for evaluating an ML model include the non-stratified, random sampling of a pre-existing population of data, as well as combining multiple datasets [140] and web-scraping *en masse* [57, 117, 130]. This is borne out in a survey of papers in the 2022 Neural Information Processing Systems (NeurIPS 2022) Datasets and Benchmarking track [1], in which we randomly sampled 20% of accepted papers, then augmented the set with the paper that received the "best paper" award for the track [130] for a total of 32 papers (details in Appendix A). Of those we reviewed, 34% of the papers build off of or combined existing datasets and 31% collected web data. Furthermore, a large minority of papers make claims on representation (12.5%) or diversity (25%).

With these techniques, metrics will be dominated by majority groups and may not accurately reflect the model's performance for those in the minority, *i.e.* the practitioner may not have visibility into when the model is poorly performing for a group with less representation in the dataset, either due to a lack of statistical significance for the minority group or due to the fact that the group's portion of the set may not contain the necessary diversity or breadth of relevant samples. This is undesirable from a robust evaluation standpoint, and also from a representation and fairness perspective [23, 47]. Considering **example 2**, if an evaluation set is created for this application by taking a random sampling of Portuguese content, *e.g.* on social media or web scraping, it is likely to be overwhelmingly dominated by content in Brazilian Portuguese due to population and volume, and be overpopulated with the most common speech. Not only is this sampling unlikely to capture the types of hateful Portuguese speech needed to evaluate the system for anyone beyond the majority group, but text scraped from the web will also be overwhelmingly from wealthier people with access to the internet, and therefore may perform less well as an evaluation for poorer populations. Moreover, less populous and lower resource varieties of Portuguese, *e.g.* African Portuguese varieties spoken in Mozambique and Angola, may not be captured sufficiently well for statistical significance, or lack important linguistic complexities such as false cognates and political speech that is contextual to these countries.

This effect is clearly shown in De Vries et al. [37], where the authors contrast the geographic distribution of content in large image datasets with that of the population of the world. They show that computer vision models trained on publicly available object-recognition systems perform relatively poorly on common household items from countries with a lower household income when a dataset that has more "representative geographical coverage" [37] is employed in evaluation [54, 121]. Their findings not only point to poor geospacial and socioeconomic representation in large image datasets, but also demonstrate the value of a representative evaluation dataset in uncovering otherwise invisible model shortcomings. Similar findings on geographic and sociocultural diversity in large datasets have been previously documented [83, 117, 123, 132]. To explore how to move beyond standard practice, we discuss approaches observed in literature that grapple with different aspects of, and techniques towards, better representation for evaluations.

3.2 Sampling by group

The first practice to integrate into our framework for instrumentally representative evaluation datasets is to make a deliberate decision on the groups to include when evaluating the model, then ensuring each group is included in the dataset, *e.g.* by stratifying sampling by data from or correlated with that group. Our NeurIPS survey demonstrates that this is not standard practice; three papers [30, 57, 86] in our NeurIPS survey explicitly consider groups in dataset construction.

We are employing "group" as a shorthand for an attribute in the data that is tightly correlated with groups of humans. These attributes may not explicitly reference data on human characteristics that belong to demographic group (*e.g.* images of faces or instances of language) but instead correspond to what one or another group might experience (*e.g.* common objects or street signs in a country). The groups in the dataset are attributes of the *data subjects*, defined in Denton et al. [41] as those "whose likenesses or utterances are documented and absorbed into the dataset." Notably, data subjects are not necessarily the *user* of the model, yet it is their features or experiences being modeled.

For **example 2**, the practitioner would integrate this practice by ensuring that when they collect samples for their evaluation, they include those from the varieties of Portuguese spoken in the regions where the model will be deployed or have an effect (*e.g.* by stratifying sampling). Examples of this practice in literature include ensuring inclusion of these groups, or even balancing the raw number of samples from each group in the dataset (*e.g.* [30, 66]). However, the amount of data is distinct from the diversity of content in each group. For example, a practitioner may sample the same amount of data from dialects of Portuguese spoken in multiple countries, but if there are difficulties in sampling for content from *e.g.* Mozambique, the set may be less representative of Mozambique's

[3] For example, false cognates are when different words or phrases can appear the same but have drastically different meanings, *e.g.* in different dialects. This happens to be the case for the term "bicha" which is a queue (line people wait in) in Portugal, but a slur in Brazil.

hate speech lexicon. This could mean the dataset contains more duplicates, is less information-rich, or simply "bland" [15, 47, 118].

Group selection. Several complex questions come to the foreground when considering groups, first-and-foremost being group selection. What are the relevant groups, and how can their relevance be determined? One approach in group selection is to choose among groupings that have become standard in the AI fairness field, *e.g.* age, gender, race, language, and geo-location (*e.g.* [4, 29, 66, 78]). These groupings are generally drawn from census categories or discrimination law [9] yet likely exclude many relevant groups, and can break down in cultural contexts outside the US, Canada, and Europe [124].

An example of this practice in the field is the Casual Conversations dataset [66]. The Casual Conversations dataset contains video clips with sound, where paid participants carry out various common actions such as waving to the camera. The authors made an effort to balance, and clearly report on, the amount of examples for each of the chosen groups (*e.g.* self-identified, binary gender) in their sampling. The initial domain of application for the dataset is DeepFake detection. However, in releasing the dataset the aim is to allow it to be employed more widely, particularly for the purpose of rooting out fairness issues in models [66]. The follow-up iteration of Casual Conversations includes a nuanced, in-depth consideration of groupings [65].

Good practice for group selection will naturally be dependent on the context and goals of the technology in question [63, 118], *e.g.* the common demographic groupings in India are different from the standard groupings based off the sociopolitical context of the United States [124]. Further, the communities included in model evaluations tends to reflect who practitioners had in mind while building the system. It may be implicit, yet these are the communities prioritised for the accrual of the benefits of the model, and limitation of harms [17, 70, 115, 118].

While the standard groupings are often not a bad place to start, we urge practitioners to look beyond them, and seek greater understanding of the data subjects, the impacted or relevant groups, and the groups that may have an intrinsic right to be included in the evaluation (*i.e.* have the right to have their experiences, views, welfare taken into consideration). An important *first* step would be to understand how groups might express differently in the domain being modeled, for example via in-depth research, subject matter and experiential expert consultation, and participatory techniques [16, 21, 31, 88, 92]. For evaluating an AI system, what is of utmost importance to consider is: who will be affected by the model? Who must this model be high performing for? There is a growing literature on impact statements and ethical foresight that can guide practitioners' efforts to directly consider relevant groups [99, 101, 102, 110].

Intersectionality. In practice, group selection will often be an iterative process [118] and can be complicated by the question of not just how to divide groups, but also how many aspects of identity and lived experience are salient. This is a challenging question as many of the domains practitioners will consider (*e.g.* language) commonly exist on a continuum, where divisions and intersectional groups can be divided with no clear ideal stopping point.

Yet, intersectionality has been shown to be powerfully important in the evaluation of AI systems, with past research highlighting how the failure to invoke this lens can lead to system failure for certain people, though the system appeared to work well when subjected to non-intersectional group analysis [23, 76, 143]. The question of group granularity cannot be "solved" so-to-speak, but rather presents the need for careful deliberation over prospective trade-offs. Data availability or the cost of obtaining high-quality labeling per-intersectional group will frequently become a limiting factor [15, 66]. Furthermore, identity should not be thought of as a static grid [84]. In this vein, Puar [111] argues that the concept of intersectionality itself runs the risk of reifying ever more granular differences that "infinitely multiply exclusion" and advocates for the notion of *assemblage* to foreground the assembling of indiscreet, continuously varying patterns of identity.

Deliberate group choices are an improvement on standard practices, however only considering the number of samples per group may not be sufficient to cover the events that a model would need to be evaluated on in order to show the system works well for the different groups. In fact, it is difficult to say what representation should look like without a clear scope or application for the evaluation, as the practitioner cannot know who will be affected by the model, *i.e.* for whom they should ensure the model is performant. Similarly, an auditor or practitioner may not have a clear indication for whom to test the model. For example, even with the scope defined in **example 2**, it may not known beforehand in which region a new app with a built-in Portuguese hate speech classifier will be the most popular, or have the most impact.

3.3 Sampling informed by domain or application

The next practice to highlight involves scoping the evaluation [55], or the *domain*, then systematically constructing the dataset to cover a diversity of events across the domain. The important relevant events can be understood via sociotechnical practices such as subject matter consultation or group participation. To be clear, this use of "domain" veers slightly from the common usage in ML and refers to a distinct region, territory, or scope covered by the evaluation (or, the scope of the model's use, which the practitioner then builds an evaluation to cover). Ideally, the domain will have some definition to it, where the more clearly and more focused it is defined, the more feasible it is to construct an evaluation that meaningfully captures the important events in the scope. It can be thought of as akin to the notion of territory of land, where the more vast and complex the land is, or the less-defined the area, the more difficult it will be to capture the full set of salient aspects in the region.[4] In **example 1**, it is more feasible to create a representative set of street signs for one country than everywhere in general. In fact, the more countries that are included, the more salient the variation in signs is likely to be, and the more difficult it becomes to create a fully, instrumentally useful representation. For an *application*, the scope is typically well defined and tied to an expected or hoped-for real world impact. An application can be a focused portion of the entire domain that is covered by a model, or it can be the entire scope of the model.

[4]Notably, pure size is not what matters here. For example, for the photo example in Section 2, a large, many-pixel, uniformly white image is easy to compress (or represent with) a one-pixel, blank white image.

Domain information would be taken into account in **example 1**, by sampling for street signs such that the set captures the diversity of different sign types. In **example 2**, the domain is Portuguese language hate speech. To systematically sample for the domain, the practitioner would need to consider the field of hate speech and Portuguese languages directly and sample accordingly to build out a set with the fuller coverage. Doing so would require an understanding of the many themes and varieties of hate speech in Portuguese – a complex topic and an area of subject-matter expertise. This practice can guard against distribution shift, when there is a lack of portability of the model from testing to operation in the wild, leading to a certain fragility of the model and lack of robustness of the evaluations. By specifying the domain and context for the dataset as clearly as possible, and building an evaluation informed by the domain of application, practitioners can avoid falling into a "portability trap" [131].[5]

This practice of scoping a domain, then researching and systematically sampling for coverage of important attributes with the domain, is exemplified by Ribeiro et al. [120] and their `CheckList` tool. `Checklist` constructs an NLP sentiment evaluation dataset with a broader diversity and coverage of sentences and sentiments. The researchers then use `CheckList` to evaluate popular NLP models and surface previously unseen bugs, showing the benefits of evaluations that capture a domain more thoroughly – and are therefore more thoroughly representative.

On the one hand, `CheckList` builds out evaluation datasets through being deeply informed by linguistics and the domain of NLP applications [120], on the other hand, the Bender Rule – that practitioners "must always name the language(s) [they are] working on" [11, 12] – is broken in the paper. It can be gathered from the paper that `CheckList` is only designed for English, yet that is not declared. And even within English there are varieties that have different linguistic properties [18], and thus would not especially be captured by `CheckList`'s evaluations. `CheckList` evaluations may be *more* representative than other NLP tests that have not been built out with such careful linguistic domain knowledge, however it is still limited as the work has not directly contended with what groups should be taken into account in building out a representative evaluation.

3.4 Subject-domain representative evaluations

As indicated, a purely group-stratified (Sections 3.2) or domain-only (Section 3.3) approach to evaluations contains gaps, potentially leading to blindspots in the practitioners' understanding of the model performance and representational harms, typically falling on historically minoritised communities. Further, attributes that belong to groups or subjects, and those that describe a domain are not always fully separable. In this section we introduce *subject-domain representation* to close those gaps, leading to a more robust evaluation methodology. Specifically, subject-domain representation combines the domain and group-centered approaches, where the dataset or evaluation is constructed such that it is inclusive of the diversity of events of who or whose situations are being modeled (the data subjects).

This approach has similarities to the concept of *user-representative* sampling introduced in the *Gender Shades* work [23, 116],[6] however is meaningfully different. First, subject-domain representation moves us away from the tendency in the tech industry to prioritise the user over the data subjects, the communities who are directly impacted by the dataset and whose features or experiences are being modeled (and thus evaluated by the dataset). For example, for a facial recognition system such as the ones described in Buolamwini and Gebru [23], the user is more likely to be a software engineer or police officer than an individual with a picture in the facial recognition database. The end-users of facial recognition systems *may not* have different demographics than the data subjects. However, attention should be oriented towards evaluating the model for communities whose images will pass through the system once in operation and who will be most impacted. As another example, for a medical AI system screening x-ray images, the user is the medical professional, yet of course the subject of the AI evaluation (and training) dataset are patients – as well as those most impacted by the model performance (and the effectiveness of the evaluations). The second divergence is that user-representativeness includes a call for "equal representation" per group *e.g.* balancing sampling per-group [116]. Subject-domain representation does not argue for this, rather for sampling such that the evaluation is inclusive of the nuance and unique expression in the domain that each group may have.

In **example 1**, a subject-domain representative set would contain an image for each type of street sign in all locations where the sign recognition system will be deployed, thus allowing the practitioner to understand the performance for the system for each group (here, country). If, say, another quality such as how the sign is lit is important for image recognition, and different countries light their signs in different ways, then subject-domain representation calls for an image of each sign type per country, lit in the manner of that country. Of course, instead of lighting it could be that image quality varies per country, etc. For the Portuguese hate speech classifier in **example 2**, subject-domain sampling will not necessarily lead to a distribution of samples that is proportional to the size of the populations that speak Portuguese. For example, let's say, *hypothetically*, that the practitioner consulted domain experts and learned that Portuguese spoken in Portugal contains ten times the number of unique hateful phrases and words as in the Brazilian-Portuguese lexicon. This indicates that achieving similarly representative coverage of hate speech that is instrumentally important to ensuring there are as few gaps as possible in the evaluation, would require more examples of language from Portugal. In this hypothetical case, for a subject-domain representative dataset, there might need to be

[5]The portability trap is defined by Selbst et al. [131] as the "failure to understand how repurposing algorithmic solutions designed for one social context may be misleading, inaccurate, or otherwise do harm when applied to a different context."

[6]User-representative sampling: "...the benchmark does not have proportional demographic distribution of the intended user population but representative inclusion of the diversity of that group. With equal representation of each distinct subgroup of the user population regardless of the percentage at which that population is present in the sample of users, we can thus evaluate for equitable model performance across subgroups" [116].

ten times the raw number[7] of unique hateful samples from Portugal in our evaluation data, even though the Portugual-Portuguese speaking population is much smaller than the Brazilian-Portuguese speaking population. Of course, the same logic would apply for other varieties of Portuguese such as those spoken in Angola and Mozambique.

With this in mind, given that the practitioner (1) identifies the groups their system should function well for (Section 3.2), and (2) has scoped the domain in which the system should be performant, the practitioner's next challenge is to sample for each group such that the group subsets are representative of the events which they are seeking to model. This approach is similar to the representative-inclusion and diversity concepts presented in Fazelpour and De-Arteaga [47]. Constructing evaluations for subject-domain representation also directly contends with what Chasalow and Levy [26] call *evaluation bias*, *e.g.* rewarding models that perform well on unrepresentative data of those who will be impacted.

The goal of subject-domain representation is not just to be inclusive for its own sake, but to be instrumentally beneficial – to ensure that the evaluation will surface as many failure modes in the model as possible prior to deployment. Buolamwini and Gebru [23] employ user-representative sampling for the *Gender Shades* audit dataset, Pilot Parliaments Benchmark (PPB), sampling public images of the faces of public figures in national parliaments the world-over. While this may not perfectly capture all the instances of different types of faces, it ensures there is global representation. The authors then investigate intersectionality (darker-female, lighter-male, darker-male, lighter-female as based on the Fitzpatrick skin tone groupings) and evaluate facial recognition systems per-group. With the framework of subject-domain representation, one can conceive of further work to augment PPB for greater representativity *e.g.* including information richness and diversity metrics (Section 5.4), and perhaps sampling based on face shapes or other important domain attributes in images, like lighting or pose. In Raji and Buolamwini [116], the authors find there is a reduced performance disparity between groups in the targeted companies' facial recognition systems after the *Gender Shades* publication. Further examples of datasets that are edging towards subject-domain representation include the ROOTs [81] and the Dollar Street datasets [54, 121].

In creating datasets that are balanced by group or for subject-domain representation, an investigation is required into which groups (1) the system must work well for and (2) where groups have meaningful difference relevant to the application. Throughout the process of AI development, relevant groups may change *e.g.* with further careful consideration, consultations with (expected) impacted communities and user studies [88]. Though the need to consider representation is often presented as a static phenomenon, with decisions on representation being made at a single moment in time, in actuality the development of representative datasets is an iterative and evolving process (Sections 5.3–5.4).

Furthermore, the approaches employed in standard practices, and sampling informed by domain do not circumvent the decision on which groups to include in the evaluation. Rather, the decision has been made implicitly and can be surfaced by analyzing which groups have been represented in the dataset. In a similar vein, sampling informed by the domain of application and subject-domain representation require ML practitioners to make a deliberate decision on the scope and goals of the system. If the practitioner would prefer not to decide, for example in an endeavor to keep the system as general as possible, a decision as to what domain the model will be performant on has still been made, just implicitly – there will still be tasks for which a general system is better or worse – rather than deliberately and transparently. By surfacing the scope, goals, and thus limitations of the system, practitioners can make intentional choices and curate an evaluation set such that the dataset, and then the system, can be robust to those goals.

4 LIMITATIONS AND TENSIONS OF REPRESENTATION IN DATASETS

Drawing upon the sociotechnical literature, it is widely understood that AI models are not merely mathematical constructs, but sociotechnical and political entities, with inherent value systems embedded in the choices made by designers about how to create and implement the model [8, 17, 60, 115, 124, 127, 134]. While there are harms that arise from improper or incomplete representation, perfect representation of all but the simplest situations is typically impossible (or infeasible) – as explained previously with the example of a picture representing the three-dimensional, infinite resolution reality, or the example presented by Pettit [106] that a fully representative democracy includes exactly everyone and is an impractical form of governance. Similarly, there cannot be a benchmark or evaluation dataset that fully represents reality, a continuous spectrum, or even fairly complex concepts (such as general perception) [70, 115]. Ultimately, not everything can or even should be represented [73].

Knowing this, what are the limits of representation? From the vantage point of political and moral philosophy the central question is: what is fair representation? And are there things that cannot be represented – or ought not to be represented, even when it is possible to do so? Examples of the former may be traits or identity types that cannot be meaningfully observed, or are essentially contested or fundamentally fluid (*e.g.* sexual orientation or gender identity) [73, 84, 137]. Examples of the latter could be traits that place people at inherent risk or that sit in tension with the notion with ideas about privacy, respect and dignity (*e.g.* [45, 149]). Indeed, increasing representation in datasets is not a panacea – it may not even be desirable to the community being represented by the data [41, 90]. There are tensions and trade-offs between better representation and other desiderata for the communities and situations that would be represented.[8] The community may be opposed to, or simply not

[7] One practice to increase the prominence of a group that is smaller, or more difficult to obtain samples for, is re-weighting or up-weighting this group thereby increasing the *effective* number of samples for that group in the set. As is known from polling practices to predict election outcomes, this practice will only get you so far. Increasing the weights effectively creates copies of the heavily weighted sample – an analogy to group tokenism. This may increase the number of events for a group but does not capture the great diversity *e.g.* of political opinions, and thus does not better capture the range, diversity, or representation of political opinions of the smaller population.

[8] Notably, under-representation in datasets is not always the problem at issue. Societally minoretised communities are in fact over-represented in some technological systems, *e.g.* predictive policing algorithms in the US over-represent African American communities. However these communities do not have meaningful representation and power in the design and evaluation of these systems [85].

be interested in, being better represented in a dataset if it comes with being surveilled [10, 22, 64, 73, 90]. Beyond those concerns, there are limits to what a representative dataset can achieve. Full representation does not guarantee system will be responsible, particularly if the system itself is not in the interest of those being represented (*e.g.* in biometrics [141]). Below, we unpack some of these considerations.

Representation and data access or existence. Needless to say, to be represented in a dataset, the data needs to exist and be accessible to the practitioner. Data on a topic or from a community may not exist for any number of reasons including a community's lack of access to – or interest in accessing – technologies that are collecting or creating the data (*e.g.* internet access) [37, 45, 132]. It could also be due to choices or resource limitations on behalf of the practitioners' to collect data specifically from particular communities (*e.g.* behavioural studies on college students). There are cases where the data doesn't exist due to the fact that the topic is fundamentally fluid, *e.g.* gender or many types of disability [137, 138]. In other cases, submitting to being tracked and their data being accessed (or even created at all) goes against the interests of the community [73]. In these situations, representation in the evaluation dataset cannot be present without the data existing and without access to it by the group implementing the evaluation. For such circumstances, alternatives need to be considered in order to ensure fair and just implementations, for example: models of auditing and evaluation that ensure true consent and sovereign ownership of data by data subjects, participation in design, and even fundamental inquiries as to whether the technology should exist [16, 31, 41, 49, 71, 137].

Representation and foundation models. Due to the fact that *foundation models* [20], or general purpose models, are meant to be separate from their eventual applications – in fact part of their goal is to be somewhat disembodied from a tightly defined domain – it is difficult to determine what a subject-domain representative evaluation (or even any comprehensive evaluations) of these models should look like.[9] This limitation echoes the arguments against widely-scoped benchmarks [115] and difficulties highlighted by previous researchers on the topic of envisioning the impact of AI applications [99, 102, 110]. That said, lack of representation in foundation models models can limit the effectiveness of downstream models built on them. For example, a language model trained only with English will be largely ineffective for other languages, and an image model trained with images only from North America might lead to worse performance for images from other parts of the world. Since other downstream models are built on top of foundation models, it is important to understand possible failure modes (which groups or domains the model might not work for) to (1) increase the variety of tasks and groups the foundation model can be useful for and (2) inform practitioners building on foundations models for use in specific applications. Additionally, details on how foundation models are trained and which data is used is frequently opaque, making the representative evaluations approach a vital tool for understanding how these models might impact users. Evaluations considering potential groups or domains could be useful early indicators of problems that could arise in foundation models and compound in downstream applications.

Representation and surveillance. To be represented in a dataset typically means being surveilled. For example, if an individual refuses a facial recognition scan, their face will not be represented in the facial recognition database, unless there exists an adequate proxy. Thus, there is a direct tension between representation of marginalised communities in datasets and understandable apprehension of surveillance. Further, it requires that what is to be represented is measurable, and should be measured [68, 73, 84, 137]. Some communities will not want their data collected under any circumstances, yet there are others who may be amenable to their data being included in particular contexts, *e.g.* when it is for an application they agree with, under strict usage circumstances or by a party they trust. These are central topics in the areas of privacy and transparency, and in the end are for the individual or communities themselves to decide [16, 21, 31, 49, 88, 92, 109]. These discussions on representation (or fairness, transparency, etc.) should not distract attention from the question as to whether or not the technology should be deployed at all [61, 75, 141].

Representation and power. As discussed, AI systems are sociotechnical and political entities that are not value neutral and therefore can perpetuate existing social power dynamics through their development and deployment. The topic of instrumental representation in data distributions does not interrogate the way the system that employs that data is functioning in society today, perpetuating or challenging the status quo. Nor does it ask what communities the system fails to reach, what features, functionality, or metric might better serve certain communities, or how to empower communities and upend systemic inequalities. Making a product or model work better, by developing a better evaluation, is still ultimately serving the interests of those who own, develop, deploy that product or model.

For evaluation data, design choices impacting representation are indicative of who practitioners envision will accrue the benefits of a system, yet they may not reflect the wishes of data subjects and communities who may be marginalised by downstream AI systems. Reflecting on the provocation in Kalluri [71], 'how is AI shifting power?' we consider how representation fits within the power dynamics of dataset development. The notions of representation discussed thus far treat what is represented as passive, something to be observed rather than individuals and communities with an active and empowered role in the construction of the system. We recognise that subject-domain representation addresses sampling and curating for evaluating an AI system, but it does not directly contend with empowerment of the communities that would be impacted, and representing the wishes of the data subjects and communities themselves. Trade-offs are inevitable *e.g.* inclusion can on occasion be violent [68], which begs the question of who is empowered to consider the trade-offs and decide? Indeed, a fuller actualization of representation necessarily includes the notions of political representation touched upon in Section 2, in which people are not simply present in the dataset but hold sovereign control [114]. In order to shift power in the right direction we need to (1) interrogate the values behind datasets [124] and (2) explore how data subjects can (re)gain control over their data, armed with the information to decide whether they wish to be represented [114].

[9] One might further argue that no general purpose AI system could truly be shown to exist, as a fully general (or representative) evaluation, or set of evaluations, cannot be created. This is an extension of the central point of Raji et al. [115], that there cannot be an "everything" benchmark.

Shifting power via representation gestures towards a typology reminiscent of Arnstein's *Ladder of Citizen Participation* [7], where forms of participation range from manipulation to citizen control. This "ladder of representation" could range from exclusion and invisibility (non-representation), objectification and tokenism (see Footnote 7), to nuanced community inclusion, sovereignty, and control. Viewed against this backdrop, representation as data in evaluation is the bare-minimum.

5 FUTURE RESEARCH

A solid grasp of the practicalities of beneficial and effective representation in evaluation is a powerful resource when it comes to promoting accountability – both for the purpose of rendering potential governmental policies more effective [46, 98], and to push against poor practices such as re-purposing datasets for use beyond their original contexts [131]. It is important to foster further research examining representation – a capacious, but powerful term that is deeply rooted in democratic histories. In this section we lay out a suite of topics that could be included on future research agendas in this space.

5.1 Representation along the AI pipeline

As mentioned in Section 1, there are a number of sites of representation or exclusion along the AI development pipeline, each of which needs to be examined. For example, in this work we focus on representation in evaluations but what is represented in the training data can impact performance across different groups at inference time. Indeed, Rolf et al. [122] empirically demonstrate fairer outcomes when representation of data is explicitly considered at training time. And, as mentioned, we caution that naively balancing datasets across various groups may not fully mitigate issues of representation in the final model [144].

Furthermore, past research indicates that the meaning of representation in annotation is highly contextual, varying in accordance with the needs and capabilities of the annotators, coupled with the sociopolitical and/or sociolinguistic aims of the technology [15, 92, 103, 108, 148]. In this case, the kind of representation needed might extend beyond diversity in hiring to include a careful consideration of capacities, survey-formulation, tooling and support systems to ensure that annotators are not, for example, burdened with ambiguous tasks and trapped in decision-trees or poor-UX interfaces that do not allow them to collaborate or skip tasks [15, 43, 58, 103]. Similar representation-related concerns have been extended to probing and red teaming for generative AI [53, 56]. Fully understanding how representation along the AI pipeline impacts performance on representative test sets, and leads to downstream societal effects, is a crucial question for practitioners.

5.2 Refocusing representation on impacted communities

In this work we present subject-domain representation as better practice when considered from a more prescriptive, instrumental point of view which focuses on increasing representation in evaluations to improve their capacity to surface issues in the model for different groups. This approach operates on a base-line assumption that the model must work sufficiently well for groups in a given context, but does not take full account of the way in which the model is likely to be used and deployed.

Ultimately, it is important to reach beyond this frame of analysis – and ensure that the AI system is particularly robust for communities that encounter specific challenges or have the least ability to absorb the harms of errors. If we consider **example 2**, Portuguese hate speech in countries with high levels of political instability could be particularly harmful. As a consequence, there may be further need to collect more and higher quality samples for these contexts, to carry out more nuanced participatory engagements, to have more careful monitoring in place, and to run additional checks on the Portuguese that is used in countries that evidence this dynamic. Outside the evaluation set, the threshold for classifier false positives or negatives may also be more stringent prior to deployment. In fact, the situation may be so acute that it becomes necessary to restrict the domain and groups only to that country, working on a case-specific basis, and potentially evaluating a model with only this at-risk group in mind, instead of aiming at a general model evaluation for all hate speech detection in Portuguese. With consideration of these questions, evaluations will get closer to forms of representation that anticipate and improve the impact that a system may have on human welfare. Given the distribution of the global population and distribution of case-specific concerns, traversing this frontier – which involves shifting from data subjects to model impact – is important in order to avoid the continuation of historical under-representation in the tech sector of communities in low- and middle-income countries.

5.3 Representative of when?

Beyond "representative of who?" and "what?" one may also need to ask, "representative of when?" This is a subtle and important question with relevance across the entire field of ML. Datasets themselves are historically situated artefacts relevant not only to domains and groups, but also to a point in time. As a consequence, who and what is being represented by the dataset (the "models of reality" [70]) will likely change over time [41]. Considering **example 1**, a dataset could be representative of the types of street signs in Indonesia in 1985, but if there is an overhaul of their signage, the dataset may not continue to be an accurate representation of the country's street signs today. This simple example does not even account for the inherent fluidity of some notions of identity that are often encoded in datasets [84]. However, it demonstrates that decisions to update or re-use a dataset have both practical and political dimensions. Given that the quality of dataset representation likely degrades over time, time itself is an important ethical consideration, and a central concern of distributional shift research [129].

5.4 How to operationalise instrumentally beneficial representative evaluations

We understand from Section 4 that there are crucial limitations to an instrumental-only approach to representation in datasets, yet it is still appropriate to explore how one can build a subject-domain representative evaluation. Here, we provide an early outline for how a practitioner (or dataset auditor) might approach this task.

(1) Design the Evaluation. Subject-domain representativeness involves developing an explicit vision of what model outcomes, or

impacts, the evaluation dataset must capture, starting with the basic questions "what is the model meant to do?" and "who will it impact? Or who has the right or need to be included in evaluation?" Subject-domain representation then requires taking this a step further, to delineate how different groups might have a different expression in a domain and thus require different sampling techniques. The varying needs of the groups may not be understandable by the practitioners on their own, given the technical training they may have and their singular lived experience [123]. Deeper understanding requires the situated knowledge of those who experience these outcomes, and best practice calls for (at minimum) the inclusion of these voices with, for example, subject matter expert consultation, participatory methods, and a staged iterative release [14, 16, 31, 88]. For example, designing a subject-domain representative dataset for **example 2** would include consulting Portuguese experts to understand the different speaker communities and map the diverse dialects, contexts, and hateful speech for each group.

(2) Measure or Audit the Dataset. Whether a dataset already exists for a task, or is to be collected and curated, the practitioner or dataset auditor will need to measure the dataset against a vision for representativity. In the case of **example 2**, the dataset could include domain-specific labels, such as length of text, parts of speech, and text topic, and group attributes, such as whether language is reflective of Portuguese spoken in Portugal, Angola, or Brazil. Representativity can be audited qualitatively, making use of data exploration tools (*e.g.* [136, 142]), or measured with metrics such as those outlined in [96] that capture properties such as diversity (*e.g.* [51, 95]) or data density (*e.g.* [32, 79]). As articulated in Mitchell et al. [96], how to effectively quantify data is an ongoing research area. We hope the sociotechnical insights herein can encourage practitioners to study and propose rigorous metrics for data representation.

(3) Build Evaluation Set. This stage involves following best practice in data collection and enrichment [103] to build out the representative dataset according to the design determined in the first step. This will include the standard practice of determining a data source, including either identifying novel data sources [37, 116, 121] or hiring annotators to generate data with a particular distribution [117]. Then, iteratively sampling based on the attributes outlined in the design stage, or measuring for those attributes in an already-sourced dataset. Given that the result will be an evaluation, it is crucial to maintain the highest standards of data quality.

(4) Monitor Representation and Outcomes. Finally, what is considered good representation is likely to change with time (Section 5.3) [41]. Additionally, distribution shift and the tendency for input data to change over time have both been documented in the broader machine learning literature [82, 151], and such shifts are likely to impact the representativeness of the data. Thus, we advocate monitoring representation over time to ensure that datasets reflect current users, data subjects, impacted communities, and values. In particular, when first designing an evaluation, practitioners could consider how often they will monitor their dataset based on the stability of the groups and domain (*e.g.* if user populations shift, every six months, or when someone raises a bug), then revisit the **Design**, **Measure** and **Build** steps as necessary.

6 CONCLUSION

In this work we focus on representation in evaluation datasets for AI models, both due to the prevalence of poor representation in datasets across the literature, and the pressing need for an understanding of representation that is aligned with the goal of creating beneficial, high-performing systems for everyone. Pulling together effective practices in the field, we provide conceptual scaffolding to construct effective and instrumentally valuable representation in model evaluations, culminating in *subject-domain representative* datasets. These datasets are sampled such that they are inclusive of the diversity of events for a carefully chosen set of groups, and informed by the application (or domain) in question.

When employed in an audit, subject-domain representation provides more reliable visibility into how the model is performing for the declared groups and domains. When coupled with transparency practices, this approach can engender trust in the performance metrics – particularly fairness and equality metrics, which might otherwise have serious limitations. For example, even high performance on fairness metrics can obscure an issue if they are based on a dataset with poor representation of any particular group.

The conceptual scaffolding of representation described in this paper has implications for the frontiers of research in AI. For example, due to their lack of tight connection to a domain of application, it is likely infeasible to design anything near an effectively representative evaluation for foundation models and notions of artificial general intelligence, given the number of domains and groups these systems are expected to cover. That said, the framework described herein can be a guide for how to envision *more* representative evaluation sets, as well as how to capably report limitations.

Even still, representation is not a panacea. There are limitations and tensions with the purely-instrumental approach presented in this paper, such as the fact that not every characteristic can or should be represented. There is more work to be done. The field of responsible AI urgently needs to clarify the powerful, capacious concept of representation, both because it is central to ensuring that the systems are beneficial and well-calibrated to the needs of those they affect, and because representation is a cherished value – that people may claim a right to – in and of itself. This is not a new story in the interdisciplinary sociotechnical and tech ethics fields. Rather, it is time for "representation" to undergo similar treatment to other terms in the AI ethics realm, in order to effectively advocate for, and deliberately build, equitable and beneficial technologies.

ACKNOWLEDGMENTS

The authors would like to sincerely thank the FAccT reviewers for their gracious feedback, as well as Shakir Mohamed and Nenad Tomašev for their detailed reviews prior to submission. Last, thank you to Miranda Bogan and Bobbie Chern for some early conversations on this idea.

REFERENCES

[1] NeurIPS 2022. 2022. Thirty Sixth Conference on Neural Information Processing Systems (NeurIPS 2022). Retrieved January 25, 2023 from https://neurips.cc/

[2] Mohsen Abbasi, Sorelle A Friedler, Carlos Scheidegger, and Suresh Venkatasubramanian. 2019. Fairness in representation: quantifying stereotyping as a representational harm. In *Proceedings of the 2019 SIAM International Conference on Data Mining (SDM)*. Society for Industrial and Applied Mathematics, 801–809.

[3] Parity AI. 2023. Retrieved January 16, 2023 from https://www.getparity.ai/

[4] V. Albiero, K. W. Bowyer, K. Vangara, and M. C. King. 2020. Does Face Recognition Accuracy Get Better With Age? Deep Face Matchers Say No. In *2020 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE Computer Society, Los Alamitos, CA, USA, 250–258. https://doi.org/10.1109/WACV45572.2020.9093357

[5] Joshua Albrecht, Abraham J Fetterman, Bryden Fogelman, Ellie Kitanidis, Bartosz Wróblewski, Nicole Seo, Michael Rosenthal, Maksis Knutins, Zachary Polizzi, James B Simon, et al. 2022. Avalon: A Benchmark for RL Generalization Using Procedurally Generated Worlds. *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks* (2022).

[6] Amit Alfassy, Assaf Arbelle, Oshri Halimi, Sivan Harary, Roei Herzig, Eli Schwartz, Rameswar Panda, Michele Dolfi, Christoph Auer, Kate Saenko, et al. 2022. FETA: Towards Specializing Foundation Models for Expert Task Applications. *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks* (2022).

[7] Sherry R. Arnstein. 1969. A Ladder of Citizen Participation. 35 (July 1969), 216–224. Issue 4.

[8] Solon Barocas, Moritz Hardt, and Arvind Narayanan. 2019. *Fairness and Machine Learning*. fairmlbook.org. http://www.fairmlbook.org.

[9] Solon Barocas and Andrew D Selbst. 2016. *Big Data's Disparate Impact*. Technical Report ID 2477899. Social Science Research Network, Rochester, NY.

[10] Alvaro Bedoya. 2014. Big data and the underground railroad. *Slate* (2014). http://www.slate.com/articles/technology/future_tense/2014/11/big_data_underground_railroad_history_says_unfettered_collection_of_data.html.

[11] Emily Bender. 2019. The #BenderRule: On Naming the Languages We Study and Why It Matters. *The Gradient* (2019). https://thegradient.pub/the-benderrule-on-naming-the-languages-we-study-and-why-it-matters/.

[12] Emily M Bender. 2011. On Achieving and Evaluating Language-Independence in NLP. *LiLT* 6 (Oct. 2011).

[13] Yoshua Bengio, Aaron C. Courville, and Pascal Vincent. 2012. Unsupervised Feature Learning and Deep Learning: A Review and New Perspectives. *CoRR* abs/1206.5538 (2012). arXiv:1206.5538 http://arxiv.org/abs/1206.5538

[14] A. Stevie Bergman, Gavin Abercrombie, Shannon Spruit, Dirk Hovy, Emily Dinan, Y-Lan Boureau, and Verena Rieser. 2022. Guiding the Release of Safer E2E Conversational AI through Value Sensitive Design. In *Proceedings of the 23rd Annual Meeting of the Special Interest Group on Discourse and Dialogue*. Association for Computational Linguistics, Edinburgh, UK, 39–52. https://aclanthology.org/2022.sigdial-1.4

[15] A. Stevie Bergman and Mona Diab. 2022. Towards Responsible Natural Language Annotation for the Varieties of Arabic. In *Findings of the Association for Computational Linguistics: ACL 2022*. Association for Computational Linguistics, Dublin, Ireland, 364–371. https://doi.org/10.18653/v1/2022.findings-acl.31

[16] Abeba Birhane, William Isaac, Vinodkumar Prabhakaran, Mark Díaz, Madeleine Clare Elish, Iason Gabriel, and Shakir Mohamed. 2022. Power to the People? Opportunities and Challenges for Participatory AI. In *Equity and Access in Algorithms, Mechanisms, and Optimization* (Arlington, VA, USA) *(EAAMO '22)*. Association for Computing Machinery, New York, NY, USA, Article 6, 8 pages. https://doi.org/10.1145/3551624.3555290

[17] Abeba Birhane, Pratyusha Kalluri, Dallas Card, William Agnew, Ravit Dotan, and Michelle Bao. 2022. The Values Encoded in Machine Learning Research. In *2022 ACM Conference on Fairness, Accountability, and Transparency* (Seoul, Republic of Korea) *(FAccT '22)*. Association for Computing Machinery, New York, NY, USA, 173–184. https://doi.org/10.1145/3531146.3533083

[18] Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (Technology) is Power: A Critical Survey of "Bias" in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Online, 5454–5476. https://doi.org/10.18653/v1/2020.acl-main.485

[19] Su Lin Blodgett, Gilsinia Lopez, Alexandra Olteanu, Robert Sim, and Hanna Wallach. 2021. Stereotyping Norwegian Salmon: An Inventory of Pitfalls in Fairness Benchmark Datasets. Association for Computational Linguistics.

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

[21] Elizabeth Bondi, Lily Xu, Diana Acosta-Navas, and Jackson A Killian. 2021. Envisioning Communities: A Participatory Approach Towards AI for Social Good. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*. Association for Computing Machinery, New York, NY, USA, 425–436.

[22] Finn Brunton and Helen Nissenbaum. 2015. Obfuscation: A user's guide for privacy and protest. *MIT Press* (2015).

[23] Joy Buolamwini and Timnit Gebru. 2018. Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. In *Proceedings of the 1st Conference on Fairness, Accountability and Transparency (Proceedings of Machine Learning Research, Vol. 81)*, Sorelle A. Friedler and Christo Wilson (Eds.). PMLR, 77–91. https://proceedings.mlr.press/v81/buolamwini18a.html

[24] Danilo Bzdok, Naomi Altman, and Martin Krzywinski. 2018. Statistics versus machine learning. *Nat. Methods* 15, 4 (April 2018), 233–234.

[25] William Cai, Ro Encarnacion, Bobbie Chern, Sam Corbett-Davies, Miranda Bogen, Stevie Bergman, and Sharad Goel. 2022. Adaptive Sampling Strategies to Construct Equitable Training Datasets. In *2022 ACM Conference on Fairness, Accountability, and Transparency* (Seoul, Republic of Korea) *(FAccT '22)*. Association for Computing Machinery, New York, NY, USA, 1467–1478. https://doi.org/10.1145/3531146.3533203

[26] Kyla Chasalow and Karen Levy. 2021. Representativeness in Statistics, Politics, and Machine Learning. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (Virtual Event, Canada) *(FAccT '21)*. Association for Computing Machinery, New York, NY, USA, 77–89.

[27] Daoyuan Chen, Dawei Gao, Weirui Kuang, Yaliang Li, and Bolin Ding. 2022. pFL-Bench: A Comprehensive Benchmark for Personalized Federated Learning. *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks* (2022).

[28] Irene Y. Chen, Fredrik D. Johansson, and David Sontag. 2018. Why is My Classifier Discriminatory?. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems* (Montréal, Canada) *(NIPS'18)*. Curran Associates Inc., Red Hook, NY, USA, 3543–3554.

[29] Xingyu Chen, Zhengxiong Li, Srirangaraj Setlur, and Wenyao Xu. 2022. Exploring racial and gender disparities in voice biometrics. *Sci. Rep.* 12, 1 (March 2022), 3723.

[30] Marta R Costa-jussà, Christine Basta, Oriol Domingo, and André Niyongabo Rubungo. 2022. Occgen: Selection of real-world multilingual parallel data balanced in gender within occupations. In *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

[31] Sasha Costanza-Chock. 2018. Design Justice: Towards an Intersectional Feminist Framework for Design Theory and Practice. (June 2018).

[32] Michael A Covington. 2009. Idea density—A potentially informative characteristic of retrieved documents. In *IEEE Southeastcon 2009*. IEEE, 201–203.

[33] Kate Crawford. 2017. The trouble with bias. (2017). https://www.youtube.com/watch?v=fMym_BKWQzk NeurIPS.

[34] Kate Crawford and Trevor Paglen. 2021. Excavating AI: the politics of images in machine learning training sets. *AI & SOCIETY* (06 2021). https://doi.org/10.1007/s00146-021-01162-8

[35] Ganqu Cui, Lifan Yuan, Bingxiang He, Yangyi Chen, Zhiyuan Liu, and Maosong Sun. 2022. A unified evaluation of textual backdoor learning: Frameworks and benchmarks. *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks* (2022).

[36] Jeffrey Dastin. 2018. Amazon scraps secret AI recruiting tool that showed bias against women. *Reuters* (2018). https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G.

[37] Terrance De Vries, Ishan Misra, Changhan Wang, and Laurens Van der Maaten. 2019. Does object recognition work for everyone?. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*. 52–59.

[38] Joseph DelPreto, Chao Liu, Yiyue Luo, Michael Foshey, Yunzhu Li, Antonio Torralba, Wojciech Matusik, and Daniela Rus. 2022. ActionSense: A Multimodal Dataset and Recording Framework for Human Activities Using Wearable Sensors in a Kitchen Environment. In *Neural Information Processing Systems (NeurIPS) Track on Datasets and Benchmarks*. https://action-sense.csail.mit.edu

[39] Chengyuan Deng, Shihang Feng, Hanchen Wang, Xitong Zhang, Peng Jin, Yinan Feng, Qili Zeng, Yinpeng Chen, and Youzuo Lin. 2022. OpenFWI: Large-Scale Multi-Structural Benchmark Datasets for Seismic Full Waveform Inversion.

Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks (2022).

[40] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. ImageNet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*. 248–255. https://doi.org/10.1109/CVPR.2009.5206848

[41] Emily Denton, Alex Hanna, Razvan Amironesei, Andrew Smart, and Hilary Nicole. 2021. On the genealogy of machine learning datasets: A critical history of ImageNet. *Big Data & Society* 8, 2 (July 2021), 20539517211035955.

[42] Mark Díaz, Razvan Amironesei, Laura Weidinger, and Iason Gabriel. 2022. Accounting for Offensive Speech as a Practice of Resistance. In *Proceedings of the Sixth Workshop on Online Abuse and Harms (WOAH)*. Association for Computational Linguistics, Seattle, Washington (Hybrid), 192–202.

[43] Mark Díaz, Ian Kivlichan, Rachel Rosen, Dylan Baker, Razvan Amironesei, Vinodkumar Prabhakaran, and Emily Denton. 2022. CrowdWorkSheets: Accounting for Individual and Collective Identities Underlying Crowdsourced Dataset Annotation. In *2022 ACM Conference on Fairness, Accountability, and Transparency* (Seoul, Republic of Korea) *(FAccT '22)*. Association for Computing Machinery, New York, NY, USA, 2342–2351.

[44] Madeleine Clare Elish. 2019. Moral Crumple Zones: Cautionary Tales in Human-Robot Interaction. *Engaging STS* 5 (March 2019), 40–60.

[45] Eubanks, Virginia. 2018. *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor.* St. Martin's Press, Inc., USA.

[46] European Commission. 2021. Proposal for a Regulation of the European Parliament and of the Council Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Cerntain Union Legislative Acts. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELLAR:e0649735-a372-11eb-9585-01aa75ed71a1.

[47] Sina Fazelpour and Maria De-Arteaga. 2022. Diversity in sociotechnical machine learning systems. *Big Data & Society* 9, 1 (Jan. 2022), 20539517221082027.

[48] Shangbin Feng, Zhaoxuan Tan, Herun Wan, Ningnan Wang, Zilong Chen, Binchi Zhang, Qinghua Zheng, Wenqian Zhang, Zhenyu Lei, Shujie Yang, et al. 2022. TwiBot-22: Towards graph-based Twitter bot detection. *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks* (2022).

[49] Sue Fletcher-Watson, Jon Adams, Kabie Brook, Tony Charman, Laura Crane, James Cusack, Susan Leekam, Damian Milton, Jeremy R Parr, and Elizabeth Pellicano. 2019. Making the future together: Shaping autism research through meaningful participation. *Autism* 23, 4 (May 2019), 943–953.

[50] Sortition Foundation. 2023. Retrieved January 16, 2023 from https://www.sortitionfoundation.org/

[51] Dan Friedman and Adji Bousso Dieng. 2022. The Vendi Score: A Diversity Evaluation Metric for Machine Learning. *arXiv preprint arXiv:2210.02410* (2022).

[52] Sidney Fussell. 2019. How an Attempt at Correcting Bias in Tech Goes Wrong. *The Atlantic* (2019). https://www.theatlantic.com/technology/archive/2019/10/google-allegedly-used-homeless-train-pixel-phone/599668/.

[53] Deep Ganguli, Liane Lovitt, Jackson Kernion, Amanda Askell, Yuntao Bai, Saurav Kadavath, Ben Mann, Ethan Perez, Nicholas Schiefer, Kamal Ndousse, Andy Jones, Sam Bowman, Anna Chen, Tom Conerly, Nova DasSarma, Dawn Drain, Nelson Elhage, Sheer El-Showk, Stanislav Fort, Zac Hatfield-Dodds, Tom Henighan, Danny Hernandez, Tristan Hume, Josh Jacobson, Scott Johnston, Shauna Kravec, Catherine Olsson, Sam Ringer, Eli Tran-Johnson, Dario Amodei, Tom Brown, Nicholas Joseph, Sam McCandlish, Chris Olah, Jared Kaplan, and Jack Clark. 2022. Red Teaming Language Models to Reduce Harms: Methods, Scaling Behaviors, and Lessons Learned. (Aug. 2022). arXiv:2209.07858 [cs.CL]

[54] GapMinder.org. 2020. Dollar Street. Retrieved January 24, 2023 from https://www.gapminder.org/dollar-street

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

[56] Amelia Glaese, Nat McAleese, Maja Trębacz, John Aslanides, Vlad Firoiu, Timo Ewalds, Maribeth Rauh, Laura Weidinger, Martin Chadwick, Phoebe Thacker, Lucy Campbell-Gillingham, Jonathan Uesato, Po-Sen Huang, Ramona Comanescu, Fan Yang, Abigail See, Sumanth Dathathri, Rory Greig, Charlie Chen, Doug Fritz, Jaume Sanchez Elias, Richard Green, Soňa Mokrá, Nicholas Fernando, Boxi Wu, Rachel Foley, Susannah Young, Iason Gabriel, William Isaac, John Mellor, Demis Hassabis, Koray Kavukcuoglu, Lisa Anne Hendricks, and Geoffrey Irving. 2022. Improving alignment of dialogue agents via targeted human judgements. (Sept. 2022). arXiv:2209.14375 [cs.LG]

[57] NC Gokul, Manideep Ladi, Sumit Negi, Prem Selvaraj, Pratyush Kumar, and Mitesh M Khapra. 2022. Addressing Resource Scarcity across Sign Languages with Multilingual Pretraining and Unified-Vocabulary Datasets. In *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

[58] M.L. Gray and S. Suri. 2019. *Ghost Work: How to Stop Silicon Valley from Building a New Global Underclass.* Houghton Mifflin Harcourt. https://books.google.com/books?id=u10-uQEACAAJ

[59] Loren Grush. 2015. Google engineer apologizes after Photos app tags two black people as gorillas. *The Verge* (2015). https://www.theverge.com/2015/7/1/8880363/google-apologizes-photos-app-tags-two-black-people-gorillas.

[60] Suchin Gururangan, Dallas Card, Sarah Dreier, Emily Gade, Leroy Wang, Zeyu Wang, Luke Zettlemoyer, and Noah A. Smith. 2022. Whose Language Counts as High Quality? Measuring Language Ideologies in Text Data Selection. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, 2562–2580. https://aclanthology.org/2022.emnlp-main.165

[61] Lelia Marie Hampton. 2021. Black Feminist Musings on Algorithmic Oppression. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (Virtual Event, Canada) *(FAccT '21)*. Association for Computing Machinery, New York, NY, USA, 1. https://doi.org/10.1145/3442188.3445929

[62] Songqiao Han, Xiyang Hu, Hailiang Huang, Mingqi Jiang, and Yue Zhao. 2022. Adbench: Anomaly detection benchmark. *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks* (2022).

[63] Alex Hanna, Emily Denton, Andrew Smart, and Jamila Smith-Loud. 2020. Towards a critical race methodology in algorithmic fairness. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (Barcelona, Spain) *(FAT* '20)*. Association for Computing Machinery, New York, NY, USA, 501–512.

[64] Nabil Hassein. 2017. Against black inclusion in facial recognition. (2017). https://digitaltalkingdrum.com/2017/08/15/against-black-inclusion-in-facial-recognition/.

[65] Caner Hazirbas, Yejin Bang, Tiezheng Yu, Parisa Assar, Bilal Porgali, Vítor Albiero, Stefan Hermanek, Jacqueline Pan, Emily McReynolds, Miranda Bogen, Pascale Fung, and Cristian Canton Ferrer. 2022. Casual Conversations v2: Designing a large consent-driven dataset to measure algorithmic bias and robustness. (Nov. 2022). arXiv:2211.05809 [cs.CV]

[66] Caner Hazirbas, Joanna Bitton, Brian Dolhansky, Jacqueline Pan, Albert Gordo, and Cristian Canton Ferrer. 2021. Towards measuring fairness in AI: The casual conversations dataset. *IEEE trans. biom. behav. identity sci.* (2021), 1–1.

[67] Iwao Hirose and Jonas Olson. 2015. *The Oxford Handbook of Value Theory.* Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199959303.001.0001

[68] Anna Lauren Hoffmann. 2021. Terms of inclusion: Data, discourse, violence. *New Media & Society* 23, 12 (Dec. 2021), 3539–3556.

[69] Rodrigo Hormazabal, Changyoung Park, Soonyoung Lee, Sehui Han, Yeonsik Jo, Jaewan Lee, Ahra Jo, Seung Hwan Kim, Jaegul Choo, Moontae Lee, et al. 2022. CEDe: A collection of expert-curated datasets with atom-level entity annotations for Optical Chemical Structure Recognition. In *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

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

[71] Pratyusha Kalluri. 2020. Don't ask if AI is good or fair, ask how it shifts power. *Nature* (2020).

[72] Matthew Kay, Cynthia Matuszek, and Sean A. Munson. 2015. Unequal Representation and Gender Stereotypes in Image Search Results for Occupations. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems* (Seoul, Republic of Korea) *(CHI '15)*. Association for Computing Machinery, New York, NY, USA, 3819–3828. https://doi.org/10.1145/2702123.2702520

[73] Os Keyes. 2019. Counting the Countless: Why data science is a profound threat for queer people. https://reallifemag.com/counting-the-countless/

[74] Yo-whan Kim, Samarth Mishra, SouYoung Jin, Rameswar Panda, Hilde Kuehne, Leonid Karlinsky, Venkatesh Saligrama, Kate Saenko, Aude Oliva, and Rogerio Feris. 2022. How Transferable are Video Representations Based on Synthetic Data? *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks* (2022).

[75] Danica Kirka. 2020. UK court says face recognition violates human rights. Retrieved January 24, 2023 from https://techxplore.com/news/2020-08-uk-court-recognition-violates-human.html

[76] Youjin Kong. 2022. Are "Intersectionally Fair" AI Algorithms Really Fair to Women of Color? A Philosophical Analysis. In *2022 ACM Conference on Fairness, Accountability, and Transparency* (Seoul, Republic of Korea) *(FAccT '22)*. Association for Computing Machinery, New York, NY, USA, 485–494.

[77] Alexander Korotin, Alexander Kolesov, and Evgeny Burnaev. 2022. Kantorovich Strikes Back! Wasserstein GANs are not Optimal Transport? *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks* (2022).

[78] K. S. Krishnapriya, Vítor Albiero, Kushal Vangara, Michael C. King, and Kevin W. Bowyer. 2020. Issues Related to Face Recognition Accuracy Varying Based on Race and Skin Tone. *IEEE Transactions on Technology and Society* 1, 1 (2020), 8–20. https://doi.org/10.1109/TTS.2020.2974996

[79] Yi-An Lai, Xuan Zhu, Yi Zhang, and Mona Diab. 2020. Diversity, Density, and Homogeneity: Quantitative Characteristic Metrics for Text Collections. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*. European Language Resources Association, Marseille, France, 1739–1746. https://aclanthology.org/2020.lrec-1.215

[80] Angela Lashbrook. 2018. AI-Driven Dermatology Could Leave Dark-Skinned Patients Behind. *The Atlantic* (2018). https://www.theatlantic.com/health/archive/2018/08/machine-learning-dermatology-skin-color/567619/.

[81] Hugo Laurençon, Lucile Saulnier, Thomas Wang, Christopher Akiki, Albert Villanova del Moral, Teven Le Scao, Leandro Von Werra, Chenghao Mou, Eduardo González Ponferrada, Huu Nguyen, Jörg Frohberg, Mario Šaško, Quentin Lhoest, Angelina McMillan-Major, Gérard Dupont, Stella Biderman, Anna Rogers, Loubna Ben allal, Francesco De Toni, Giada Pistilli, Olivier Nguyen, Somaieh Nikpoor, Maraim Masoud, Pierre Colombo, Javier de la Rosa, Paulo Villegas, Tristan Thrush, Shayne Longpre, Sebastian Nagel, Leon Weber, Manuel Romero Muñoz, Jian Zhu, Daniel Van Strien, Zaid Alyafeai, Khalid Almubarak, Vu Minh Chien, Itziar Gonzalez-Dios, Aitor Soroa, Kyle Lo, Manan Dey, Pedro Ortiz Suarez, Aaron Gokaslan, Shamik Bose, David Ifeoluwa Adelani, Long Phan, Hieu Tran, Ian Yu, Suhas Pai, Jenny Chim, Violette Lepercq, Suzana Ilic, Margaret Mitchell, Sasha Luccioni, and Yacine Jernite. 2022. The BigScience ROOTS Corpus: A 1.6TB Composite Multilingual Dataset. In *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*. https://openreview.net/forum?id=UoEw6KigkUn

[82] Angeliki Lazaridou, Adhi Kuncoro, Elena Gribovskaya, Devang Agrawal, Adam Liska, Tayfun Terzi, Mai Gimenez, Cyprien de Masson d'Autume, Tomas Kocisky, Sebastian Ruder, et al. 2021. Mind the gap: Assessing temporal generalization in neural language models. *Advances in Neural Information Processing Systems* 34 (2021), 29348–29363.

[83] Fangyu Liu, Emanuele Bugliarello, Edoardo Maria Ponti, Siva Reddy, Nigel Collier, and Desmond Elliott. 2021. Visually Grounded Reasoning across Languages and Cultures. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, 10467–10485. https://doi.org/10.18653/v1/2021.emnlp-main.818

[84] Christina Lu, Jackie Kay, and Kevin McKee. 2022. Subverting Machines, Fluctuating Identities: Re-Learning Human Categorization. In *2022 ACM Conference on Fairness, Accountability, and Transparency* (Seoul, Republic of Korea) *(FAccT '22)*. Association for Computing Machinery, New York, NY, USA, 1005–1015. https://doi.org/10.1145/3531146.3533161

[85] Kristian Lum and William Isaac. 2016. To predict and serve? *Signif. (Oxf.)* 13, 5 (Oct. 2016), 14–19.

[86] Zelun Luo, Zane Durante, Linden Li, Wanze Xie, Ruochen Liu, Emily Jin, Zhuoyi Huang, Lun Yu Li, Jiajun Wu, Juan Carlos Niebles, et al. 2022. MOMA-LRG: Language-Refined Graphs for Multi-Object Multi-Actor Activity Parsing. In *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

[87] Kim Lyons. 2021. Facebook says its AI mislabeling a video of Black men as "primates" was "unacceptable". *The Verge* (2021). https://www.theverge.com/2021/9/4/22657026/facebook-mislabeling-video-black-men-primates-algorithm.

[88] Donald Martin, Jr, Vinodkumar Prabhakaran, Jill Kuhlberg, Andrew Smart, and William S Isaac. 2020. Participatory Problem Formulation for Fairer Machine Learning Through Community Based System Dynamics. ICLR Machine Learning In Real Life (ML-IRL) Workshop, Addis Ababa, Ethiopia.

[89] Mantas Mazeika, Eric Tang, Andy Zou, Steven Basart, Jun Shern Chan, Dawn Song, David Forsyth, Jacob Steinhardt, and Dan Hendrycks. 2022. How Would The Viewer Feel? Estimating Wellbeing From Video Scenarios. *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks* (2022).

[90] MediaWell. 2019. "Please do not include us": Workshop on AI Ethics and Inclusion. Retrieved January 19, 2023 from https://mediawell.ssrc.org/event/please-do-not-include-us-workshop-on-ai-ethics-and-inclusion/

[91] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2021. A Survey on Bias and Fairness in Machine Learning. *ACM Comput. Surv.* 54, 6 (July 2021), 1–35.

[92] Milagros Miceli, Tianling Yang, Adriana Alvarado Garcia, Julian Posada, Sonja Mei Wang, Marc Pohl, and Alex Hanna. 2022. Documenting Data Production Processes: A Participatory Approach for Data Work. *Proc. ACM Hum.-Comput. Interact.* 6, CSCW2, Article 510 (nov 2022), 34 pages. https://doi.org/10.1145/3555623

[93] Marvin Minsky. 2006. *The Emotion Machine*. Simon & Schuster, New York.

[94] Shubhanshu Mishra, Aman Saini, Raheleh Makki, Sneha Mehta, Aria Haghighi, and Ali Mollahosseini. 2022. TweetNERD–End to End Entity Linking Benchmark for Tweets. *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks* (2022).

[95] Margaret Mitchell, Dylan Baker, Nyalleng Moorosi, Emily Denton, Ben Hutchinson, Alex Hanna, Timnit Gebru, and Jamie Morgenstern. 2020. Diversity and inclusion metrics in subset selection. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*. 117–123.

[96] Margaret Mitchell, Alexandra Sasha Luccioni, Nathan Lambert, Marissa Gerchick, Angelina McMillan-Major, Ezinwanne Ozoani, Nazneen Rajani, Tristan Thrush, Yacine Jernite, and Douwe Kiela. 2022. Measuring Data. (Dec. 2022). arXiv:2212.05129 [cs.AI]

[97] Mazda Moayeri, Sahil Singla, and Soheil Feizi. 2022. Hard imagenet: Segmentations for objects with strong spurious cues. In *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

[98] National Institute of Standards and Technology (NIST). 2022. AI Risk Management Framework, 2nd Draft. https://www.nist.gov/itl/ai-risk-management-framework.

[99] Aviv Ovadya and Jess Whittlestone. 2019. Reducing malicious use of synthetic media research: Considerations and potential release practices for machine learning. *arXiv preprint arXiv:1907.11274* (2019).

[100] Feng Pan, Wei Wang, A K H Tung, and Jiong Yang. 2005. Finding representative set from massive data. In *Fifth IEEE International Conference on Data Mining (ICDM'05)*. 8 pp.–.

[101] Partnership on AI. 2020. Publication Norms for Responsible AI: Ongoing Initiative.

[102] Partnership on AI. 2021. Managing the Risks of AI Research: Six Recommendations for Responsible Publication.

[103] Partnership on AI. 2021. Responsible Sourcing of Data Enrichment Services.

[104] Frank Pasquale and Gianclaudio Malgieri. 2021. If You Don't Trust A.I. Yet, You're Not Wrong. *The New York Times* (2021). https://www.nytimes.com/2021/07/30/opinion/artificial-intelligence-european-union.html.

[105] Amandalynne Paullada, Inioluwa Deborah Raji, Emily M Bender, Emily Denton, and Alex Hanna. 2020. Data and its (dis)contents: A survey of dataset development and use in machine learning research. (Dec. 2020). arXiv:2012.05345 [cs.LG]

[106] Philip Pettit. 2010. *Varieties of public representation*. Cambridge University Press, 61–89. https://doi.org/10.1017/CBO9780511813146.005

[107] Hanna Fenichel Pitkin. 1967. *The Concept of Representation*. University of California Press, Berkeley. https://doi.org/doi:10.1525/9780520340503

[108] Julian Posada. 2021. Family Units. https://logicmag.io/beacons/family-units/. Accessed: 2022-11-28.

[109] Vinodkumar Prabhakaran and Donald Martin, Jr. 2020. Participatory Machine Learning Using Community-Based System Dynamics. *Health Hum. Rights* 22, 2 (Dec. 2020), 71–74.

[110] Carina Prunkl, Carolyn Ashurst, Markus Anderljung, Helena Webb, Jan Leike, and Allan Dafoe. 2021. Institutionalizing ethics in AI through broader impact requirements. *Nature Machine Intelligence* (2021).

[111] Jasbir K Puar. 2020. "I would rather be a cyborg than a goddess": Becoming-intersectional in assemblage theory. In *Feminist Theory Reader*. Routledge, 405–415.

[112] Rongjun Qin, Songyi Gao, Xingyuan Zhang, Zhen Xu, Shengkai Huang, Zewen Li, Weinan Zhang, and Yang Yu. 2022. NeoRL: A near real-world benchmark for offline reinforcement learning. *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks* (2022).

[113] Yijian Qin, Ziwei Zhang, Xin Wang, Zeyang Zhang, and Wenwu Zhu. 2022. NAS-Bench-Graph: Benchmarking Graph Neural Architecture Search. *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks* (2022).

[114] Stephanie Carroll Rainie. 2019. Indigenous Data Sovereignty. In *The State of Open Data: Histories and Horizons*. Cape Town and Ottawa: African Minds and International Development Research Centre. https://www.stateofopendata.od4d.net/chapters/issues/indigenous-data.html

[115] Deborah Raji, Emily Denton, Emily M. Bender, Alex Hanna, and Amandalynne Paullada. 2021. AI and the Everything in the Whole Wide World Benchmark. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, J. Vanschoren and S. Yeung (Eds.), Vol. 1. https://datasets-benchmarks-proceedings.neurips.cc/paper/2021/file/084b6fbb10729ed4da8c3d3f5a3ae7c9-Paper-round2.pdf

[116] Inioluwa Deborah Raji and Joy Buolamwini. 2019. Actionable Auditing: Investigating the Impact of Publicly Naming Biased Performance Results of Commercial AI Products. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society* (Honolulu, HI, USA) *(AIES '19)*. Association for Computing Machinery, New York, NY, USA, 429–435.

[117] Vikram V Ramaswamy, Sing Yu Lin, Dora Zhao, Aaron B Adcock, Laurens van der Maaten, Deepti Ghadiyaram, and Olga Russakovsky. 2023. Beyond web-scraping: Crowd-sourcing a geographically diverse image dataset. *arXiv preprint arXiv:2301.02560* (2023).

[118] Maribeth Rauh, John Mellor, Jonathan Uesato, Po-Sen Huang, Johannes Welbl, Laura Weidinger, Sumanth Dathathri, Amelia Glaese, Geoffrey Irving, Iason Gabriel, William Isaac, and Lisa Anne Hendricks. 2022. Characteristics of Harmful Text: Towards Rigorous Benchmarking of Language Models. In *Thirty-sixth*

Conference on Neural Information Processing Systems Datasets and Benchmarks Track.

[119] The Oxford Reference. 2023. *Overview: representation.* Retrieved January 23, 2023 from https://www.oxfordreference.com/display/10.1093/oi/authority.20111014165925770;jsessionid=8033B4F2345255BA0941E2E13889BD54

[120] Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. 2020. Beyond Accuracy: Behavioral Testing of NLP Models with CheckList. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Online, 4902–4912.

[121] William A Gaviria Rojas, Sudnya Diamos, Keertan Ranjan Kini, David Kanter, Vijay Janapa Reddi, and Cody Coleman. 2022. The Dollar Street Dataset: Images Representing the Geographic and Socioeconomic Diversity of the World. In *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*. 201–203.

[122] Esther Rolf, Theodora T. Worledge, Benjamin Recht, and Michael I. Jordan. 2021. Representation Matters: Assessing the Importance of Subgroup Allocations in Training Data. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event (Proceedings of Machine Learning Research, Vol. 139)*, Marina Meila and Tong Zhang (Eds.). PMLR, 9040–9051.

[123] Nithya Sambasivan. 2021. Seeing like a dataset from the global south. *Interactions* 28, 4 (July 2021), 76–78.

[124] Nithya Sambasivan, Erin Arnesen, Ben Hutchinson, Tulsee Doshi, and Vinodkumar Prabhakaran. 2021. Re-Imagining Algorithmic Fairness in India and Beyond. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (Virtual Event, Canada) *(FAccT '21)*. Association for Computing Machinery, New York, NY, USA, 315–328. https://doi.org/10.1145/3442188.3445896

[125] Nithya Sambasivan, Shivani Kapania, Hannah Highfill, Diana Akrong, Praveen Paritosh, and Lora M Aroyo. 2021. "Everyone wants to do the model work, not the data work": Data Cascades in High-Stakes AI. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21, Article 39)*. Association for Computing Machinery, New York, NY, USA, 1–15.

[126] Kate Sanders, Reno Kriz, Anqi Liu, and Benjamin Van Durme. 2022. Ambiguous Images With Human Judgments for Robust Visual Event Classification. *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks* (2022).

[127] Morgan Klaus Scheuerman, Alex Hanna, and Emily Denton. 2021. Do Datasets Have Politics? Disciplinary Values in Computer Vision Dataset Development. *Proc. ACM Hum. Comput. Interact.* 5, CSCW2 (2021), 317:1–317:37. https://doi.org/10.1145/3476058

[128] Lars Schmarje, Vasco Grossmann, Claudius Zelenka, Sabine Dippel, Rainer Kiko, Mariusz Oszust, Matti Pastell, Jenny Stracke, Anna Valros, Nina Volkmann, et al. 2022. Is one annotation enough? A data-centric image classification benchmark for noisy and ambiguous label estimation. *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks* (2022).

[129] Jessica Schrouff, Natalie Harris, Oluwasanmi O Koyejo, Ibrahim Alabdulmohsin, Eva Schnider, Krista Opsahl-Ong, Alexander Brown, Subhrajit Roy, Diana Mincu, Chrsitina Chen, Awa Dieng, Yuan Liu, Vivek Natarajan, Alan Karthikesalingam, Katherine A Heller, Silvia Chiappa, and Alexander D'Amour. 2022. Diagnosing failures of fairness transfer across distribution shift in real-world medical settings. In *Advances in Neural Information Processing Systems*, Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho (Eds.).

[130] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. 2022. Laion-5b: An open large-scale dataset for training next generation image-text models. *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks* (2022).

[131] Andrew D. Selbst, Danah Boyd, Sorelle A. Friedler, Suresh Venkatasubramanian, and Janet Vertesi. 2019. Fairness and Abstraction in Sociotechnical Systems. In *Proceedings of the Conference on Fairness, Accountability, and Transparency* (Atlanta, GA, USA) *(FAT* '19)*. Association for Computing Machinery, New York, NY, USA, 59–68. https://doi.org/10.1145/3287560.3287598

[132] Shreya Shankar, Yoni Halpern, Eric Breck, James Atwood, Jimbo Wilson, and D Sculley. 2017. No Classification without Representation: Assessing Geodiversity Issues in Open Data Sets for the Developing World. NIPS 2017: Workshop on Machine Learning for the Developing World.

[133] Renee Shelby, Shalaleh Rismani, Kathryn Henne, Ajung Moon, Negar Rostamzadeh, Paul Nicholas, N'mah Yilla, Jess Gallegos, Andrew Smart, Emilio Garcia, and Gurleen Virk. 2022. Sociotechnical Harms: Scoping a Taxonomy for Harm Reduction. (Oct. 2022). arXiv:2210.05791 [cs.HC]

[134] Harini Suresh and John Guttag. 2021. A Framework for Understanding Sources of Harm throughout the Machine Learning Life Cycle. In *Equity and Access in Algorithms, Mechanisms, and Optimization* (–, NY, USA) *(EAAMO '21)*. Association for Computing Machinery, New York, NY, USA, Article 17, 9 pages. https://doi.org/10.1145/3465416.3483305

[135] Alex Tamkin, Gaurab Banerjee, Mohamed Owda, Vincent Liu, Shashank Rammoorthy, and Noah Goodman. 2022. DABS 2.0: Improved Datasets and Algorithms for Universal Self-Supervision. In *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

[136] Ian Tenney, James Wexler, Jasmijn Bastings, Tolga Bolukbasi, Andy Coenen, Sebastian Gehrmann, Ellen Jiang, Mahima Pushkarna, Carey Radebaugh, Emily Reif, and Ann Yuan. 2020. The Language Interpretability Tool: Extensible, Interactive Visualizations and Analysis for NLP Models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Association for Computational Linguistics, Online, 107–118. https://doi.org/10.18653/v1/2020.emnlp-demos.15

[137] Nenad Tomasev, Kevin R McKee, Jackie Kay, and Shakir Mohamed. 2021. Fairness for Unobserved Characteristics: Insights from Technological Impacts on Queer Communities. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society* (Virtual Event, USA) *(AIES '21)*. Association for Computing Machinery, New York, NY, USA, 254–265.

[138] Shari Trewin. 2018. AI Fairness for People With Disabilities: Point of View.

[139] Renbo Tu, Nicholas Roberts, Mikhail Khodak, Junhong Shen, Frederic Sala, and Ameet Talwalkar. 2022. NAS-bench-360: Benchmarking neural architecture search on diverse tasks. *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks* (2022).

[140] Ihsan Ullah, Dustin Carrión-Ojeda, Sergio Escalera, Isabelle M Guyon, Mike Huisman, Felix Mohr, Jan N van Rijn, Haozhe Sun, Joaquin Vanschoren, and Phan Anh Vu. 2022. Meta-Album: Multi-domain Meta-Dataset for Few-Shot Image Classification. In *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

[141] Ana Valdivia, Júlia Corbera Serrajòrdia, and Aneta Swianiewicz. 2022. There is an elephant in the room: towards a critique on the use of fairness in biometrics. *AI and Ethics* (Dec. 2022).

[142] Angelina Wang, Alexander Liu, Ryan Zhang, Anat Kleiman, Leslie Kim, Dora Zhao, Iroha Shirai, Arvind Narayanan, and Olga Russakovsky. 2022. REVISE: A tool for measuring and mitigating bias in visual datasets. *International Journal of Computer Vision* (2022), 1–21.

[143] Angelina Wang, Vikram V Ramaswamy, and Olga Russakovsky. 2022. Towards Intersectionality in Machine Learning: Including More Identities, Handling Underrepresentation, and Performing Evaluation. In *2022 ACM Conference on Fairness, Accountability, and Transparency* (Seoul, Republic of Korea) *(FAccT '22)*. Association for Computing Machinery, New York, NY, USA, 336–349. https://doi.org/10.1145/3531146.3533101

[144] Tianlu Wang, Jieyu Zhao, Mark Yatskar, Kai-Wei Chang, and Vicente Ordonez. 2019. Balanced datasets are not enough: Estimating and mitigating gender bias in deep image representations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 5310–5319.

[145] Emily Wenger, Roma Bhattacharjee, Arjun Nitin Bhagoji, Josephine Passananti, Emilio Andere, Haitao Zheng, and Ben Zhao. 2022. Finding Naturally Occurring Physical Backdoors in Image Datasets. In *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

[146] Wikipedia. 2023. Comparison of MUTCD-influenced traffic signs. Retrieved February 6, 2023 from https://en.wikipedia.org/wiki/Comparison_of_MUTCD-influenced_traffic_signs

[147] Wiktionary. 2023. *Represent.* Retrieved January 23, 2023 from https://en.wiktionary.org/wiki/represent

[148] Adrienne Williams. 2022. The exploited labor behind artificial intelligence. (Oct. 2022).

[149] Jonathan Wolff. 2010. Fairness, respect and the egalitarian ethos revisited. *J. Ethics* 14, 3-4 (Dec. 2010), 335–350.

[150] Minghao Xu, Zuobai Zhang, Jiarui Lu, Zhaocheng Zhu, Yangtian Zhang, Chang Ma, Runcheng Liu, and Jian Tang. 2022. Peer: a comprehensive and multi-task benchmark for protein sequence understanding. *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks* (2022).

[151] Huaxiu Yao, Caroline Choi, Bochuan Cao, Yoonho Lee, Pang Wei Koh, and Chelsea Finn. 2022. Wild-time: A benchmark of in-the-wild distribution shift over time. *arXiv preprint arXiv:2211.14238* (2022).

[152] Iris Marion Young. 2002. *Inclusion and Democracy*. Oxford University Press. https://doi.org/10.1093/0198297556.003.0001

[153] Chao Yu, Akash Velu, Eugene Vinitsky, Yu Wang, Alexandre Bayen, and Yi Wu. 2022. The surprising effectiveness of ppo in cooperative, multi-agent games. *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks* (2022).

[154] Klim Zaporojets, Lucie-Aimée Kaffee, Thomas Demeester, Chris Develder, and I Augenstein. 2022. TempEL: Linking dynamically evolving and newly emerging entities. In *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

[155] Kaizhi Zheng, Xiaotong Chen, Odest Chadwicke Jenkins, and Xin Eric Wang. 2022. Vlmbench: A compositional benchmark for vision-and-language manipulation. *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks* (2022).

A SURVEY OF NEURIPS 2022 DATASETS AND BENCHMARKS

To build our intuition on standard practice in the field, we performed a study on papers accepted to the 2022 NeurIPS Benchmarks and Datasets Track. In particular, we considered the best paper from the track and randomly selected 20% of the papers from the track (for a total of 32 papers). We sampled papers regardless of whether they made any claims on representativity in their paper. Two randomly sampled papers [35, 153] did not present a new dataset, but rather a testing protocol meant to be used with existing data [35] or baseline evaluations on existing data [153]. Of the rest, 12.5% were collected for a specific application. For example [48], was explicitly collected to benchmark models designed to detect TwitterBots. As opposed to specific applications, the majority of datasets were collected for existing domains of interest to the machine learning community, like object recognition [97] or anomaly detection [62]. Upon reading the papers, we found that though many datasets included domain expertise to inform their data collection (e.g., [97] builds on extensive work on common correlations in image datasets), the exact procedure for sampling data that covers a domain is frequently unclear. One paper we believe was close to domain-informed sampling is [94] which considered phrase entropy to collect tweets with diversity entity popularity and disambiguation difficulty. Only 6% of papers attempted to sample based on groups. This demonstrates that while there are calls for more representative data, practically communities building benchmark datasets are not carefully considering impacted groups.

We list here all the papers we considered in our analysis emphasising that our main aim is to understand standard practice, rather than critique any individual paper: Laurençon et al. [81], Schuhmann et al. [130], Moayeri et al. [97], Cui et al. [35], Ullah et al. [140], Gokul et al. [57], Albrecht et al. [5], Mishra et al. [94], Qin et al. [112], Zaporojets et al. [154], Deng et al. [39], Korotin et al. [77], Xu et al. [150], Yu et al. [153], Yao et al. [151], Alfassy et al. [6], Zheng et al. [155], Sanders et al. [126], DelPreto et al. [38], Kim et al. [74], Wenger et al. [145], Mazeika et al. [89], [62], Tu et al. [139], Tamkin et al. [135], Hormazabal et al. [69], Feng et al. [48], Luo et al. [86], Costa-jussà et al. [30], Qin et al. [113], Schmarje et al. [128], and Chen et al. [27].