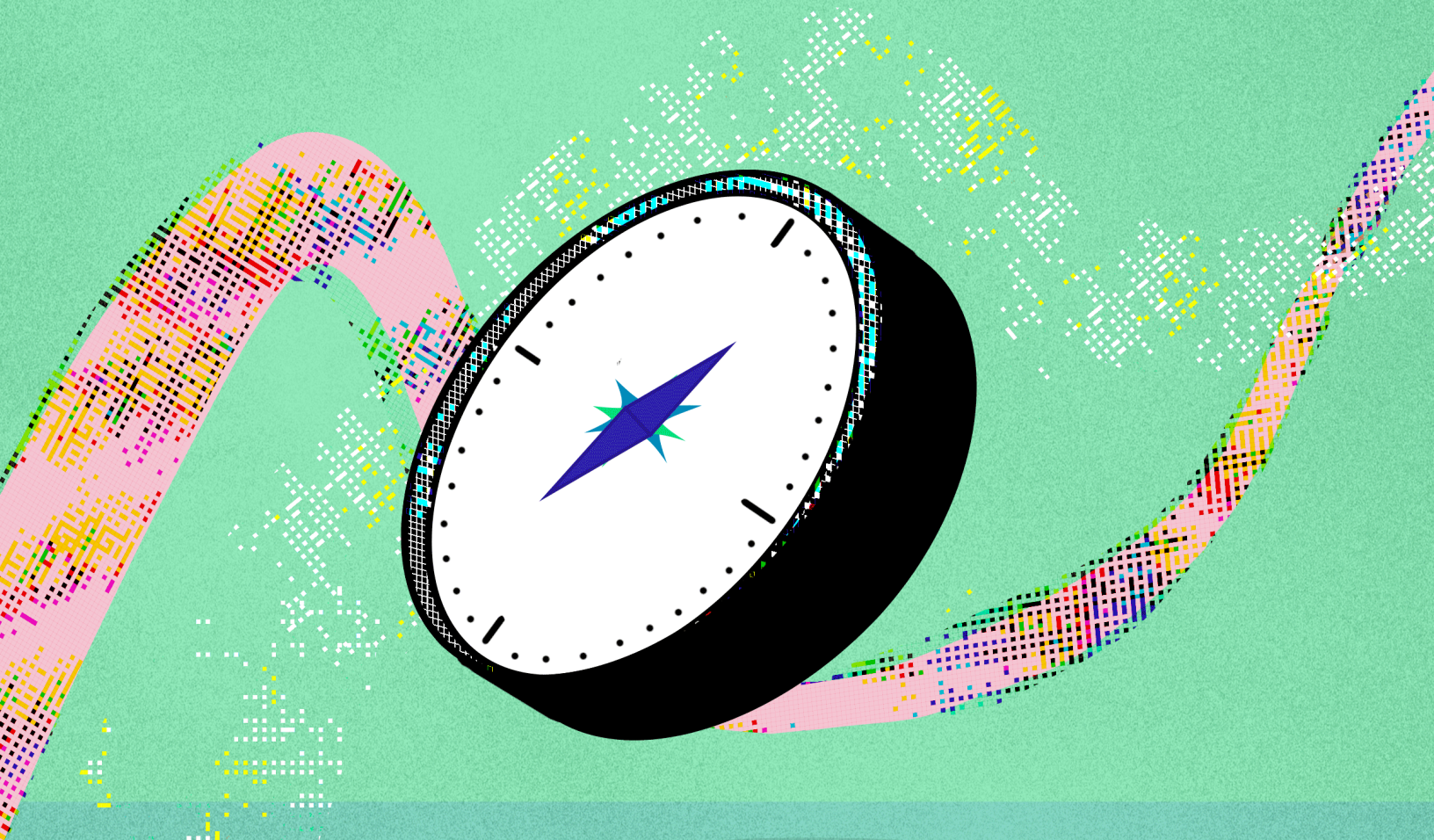


MAY 2024

Navigating Demographic Measurement for Fairness and Equity

AI Governance in Practice Guide



Miranda Bogen



The **Center for Democracy & Technology (CDT)** is the leading nonpartisan, nonprofit organization fighting to advance civil rights and civil liberties in the digital age. We shape technology policy, governance, and design with a focus on equity and democratic values. Established in 1994, CDT has been a trusted advocate for digital rights since the earliest days of the internet. The organization is headquartered in Washington, D.C. and has a Europe Office in Brussels, Belgium.



This report is licensed under a [Creative Commons Attribution 4.0 International License](https://creativecommons.org/licenses/by/4.0/).

Navigating Demographic Measurement for Fairness and Equity

AI Governance in Practice Guide

Miranda Bogen

With contributions from Ariana Aboulafia, Kevin Bankston, Ridhi Shetty and Amy Winecoff. Designed and illustrated by Timothy Hoagland.

Thanks to Quinn Anex-Ries, Gemma Galdon Clavell, Janet Haven, Stephen Hayes Ming Hsu, Logan Koepke, Laura MacCleery, Eliza McCullough, Serena Oduro, and Claudia Ruiz for their feedback.



Executive Summary

Governments and policymakers increasingly expect practitioners developing and using AI systems in both consumer and public sector settings to proactively identify and address bias or discrimination that those AI systems may reflect or amplify. Central to this effort is the complex and sensitive task of obtaining demographic data to measure fairness and bias within and surrounding these systems. This report provides methodologies, guidance, and case studies for those undertaking fairness and equity assessments — from approaches that involve more direct access to data to ones that don't expand data collection.

Practitioners are guided through the first phases of demographic *measurement* efforts, including determining the relevant lens of analysis, selecting what demographic characteristics to consider, and navigating how to hone in on relevant sub-communities. The report then delves into several approaches to uncover demographic patterns.

Given long histories of demographic data being misused to the detriment of vulnerable communities, the report emphasizes that responsibly *handling* demographic data is just as critical as the measurement methods themselves. Many of the approaches described have the potential to be mixed and matched with one another to strengthen protections against potential harms while helping to enable critical work.



Approaches for Measuring Demographic Characteristics for Fairness Measurement

Measuring disparities related to real people

- » **Collection:** Directly asking individuals to self-report their demographic information.
- » **Observation and inference:** Assigning perceived demographic characteristics based on observable features or predicting them using statistical methods or machine learning.
- » **Proxies and surrogate characteristics:** Using signals that correlate with demographic characteristics to detect patterns or disparities without directly inferring individual demographics.
- » **Auxiliary datasets:** Combining existing datasets containing demographic information with the data of interest.
- » **Cohort discovery:** Using pattern detection techniques to identify groups experiencing negative outcomes, without explicitly naming demographic characteristics.

Measuring disparities related to representations

- » **Keywords and terms:** Manually or automatically constructing lists of words and topics that relate to demographic characteristics and using them to probe systems.
- » **Observation and labeling of content:** Automatically or manually assigning labels of apparent traits to unidentified people represented in audiovisual or text content.

Measuring disparities across contexts

- » **Synthetic data:** Using artificially generated data that simulates the structure and distribution of real-world examples or populations.
- » **Exploratory analysis:** Reviewing a system to reason about how its design, behavior, or other characteristics might lead to negative impact for certain communities.
- » **Qualitative research:** Directly engaging with people using and affected by systems to capture more nuanced insights about people's lived experience.



Table 1. Approaches for Measuring Demographic Characteristics for Fairness Measurement.

Approaches for Handling Demographic Characteristics for Fairness Measurement

Data and infrastructure controls

- » **Pseudonymization:** Replaces personal identifiers with placeholder information or otherwise breaks the link between identifying data and other data about an individual.
- » **Infrastructure controls:** Data and system architecture choices that limit how and by whom data and measurement methods can be accessed or used.
- » **Encryption:** Scrambling data so it can't be easily deciphered without a mathematical key.
- » **Retention and ephemerality:** Preventing data from being created or stored longer than needed.

Privacy enhancing methods

- » **Aggregation:** Combining and summarizing data to reduce identifiability of individual data points.
- » **Differential privacy:** Adding a specific amount of random statistical noise to datasets to realize particular privacy constraints.
- » **Secure multi-party computation:** A cryptographic protocol that allows parties to conduct analyses across multiple datasets without sharing data with one another.

Procedural controls

- » **User controls:** Providing people with the opportunity to decide whether to share their data and to request data be corrected or deleted.
- » **Organizational oversight:** Processes to review proposed uses of data or measurement methods to ensure they comply with policies and follow necessary procedures.
- » **Separate teams:** Assigning a specific team to be responsible for oversight and compliance with laws that implicate demographic measurement.
- » **Privacy impact assessments:** Structured impact assessments to evaluate whether proposed use of data sufficiently mitigate against privacy risks.



Table 2. Approaches for Handling Demographic Characteristics for Fairness Measurement.

As policymakers and practitioners build regulatory and technical infrastructure to make progress in this domain, we highlight several recommendations to ensure that the balance remains tipped toward beneficial measurement efforts.

Practitioners should:

- Establish ongoing relationships with communities affected by measurement activities to co-design data collection and handling strategies, discuss potential risks and benefits, and collaboratively define fairness goals.
- Where possible, consider methods that avoid generating or storing sensitive demographic information in a way that can be easily connected to individuals.
- Take great care before using observation and inference methods to identify characteristics, especially those lacking precedent or that resist observation.
- Clearly differentiate between perceived or implied characteristics and actual ones
- Employ a robust combination of approaches to handling data and measurement methods to ensure appropriate use.
- Communicate openly about demographic measurement efforts, as well as how data is handled.

Government agencies and regulators should:

- Recognize that a variety of approaches are available for companies to identify and measure disparities, even in the absence of comprehensive demographic data collection.
- Clarify criteria and expectations about acceptable measurement methods when it comes to civil rights compliance, and articulate minimum expectations for how data and methods should be handled.
- Explore how more measurement methods can be used to monitor compliance with Federal civil rights laws, including to conduct investigations and enforcement actions.

- Facilitate collaboration between NGOs, research institutes, and government data agencies to explore creative ways that existing administrative data can be used to conduct measurements in a privacy-respecting manner.
- Encourage continued research to explore how unsupervised, synthetic, privacy-enhancing, and content-related methods can be used to further the detection and remediation of bias and discrimination.

While there is no one-size-fits-all solution, this report makes clear that the lack of obvious access to raw demographic data should not be considered an insurmountable barrier to assessing AI systems for fairness, nor should it provide a blanket justification for widespread or incautious data collection efforts. From exploring privacy-preserving techniques to pursuing measurement of content-related bias when disparities affecting people are hard to measure directly, practitioners have a range of tools at their disposal. As practitioners navigate this complex but important landscape, they should engage early and often with impacted communities, clearly document and communicate their practices, and embed strong technical and institutional safeguards to prevent misuse. Ultimately, responsible demographic measurement demands extraordinary care — for technical choices and their implications, but even more for the people and communities this work must ultimately serve.

“Respecting people’s self-determination and autonomy when it comes to sensitive data about who we are is complex and hard to do well. But ignoring that kind of data is also not an option.”¹



Contents

Executive Summary	4
Contents	9
Introduction	11
Determining the Relevant Lens of Analysis	15
Understanding and Defining Relevant Dimensions for Measurement	15
Selecting Demographic Dimensions and Segmentations	19
Approaches for Measuring Demographic Characteristics for Fairness Measurement	24
Measuring Disparities Related to Real People	25
Measuring Disparities Related to Representations	48
Measuring Disparities Across Contexts	55
Approaches for Handling Demographic Characteristics for Fairness Measurement	64
Data and infrastructure controls	65
Privacy enhancing methods	73
Procedural controls	79



Contents

Discussion and Recommendations	87
Conclusion	91
Glossary	93
Endnotes	100



01

Introduction

Governments and policymakers increasingly expect practitioners developing and using AI systems in both consumer and public sector settings to proactively identify and address bias or discrimination that those AI systems may reflect or amplify.² Civil rights and equality laws that protect people from discrimination based on characteristics like age, gender, race, disability, religion, and others (often called **protected characteristics** or “protected classes”) tend to involve significant measurement efforts, and many obvious failure modes of AI systems that negatively affect communities can be spotted through proactive analysis. In most cases, having access to demographic data of some kind is necessary to do so, since those efforts rely on awareness of how circumstances or outcomes differ across groups and how efforts to mitigate gaps affect those observations. While certain domains have well-established practices for obtaining this data, with techniques ranging from direct collection to estimation, a large proportion of public and private organizations still face

challenges or perceive barriers to bias measurement when demographic data is unavailable or incomplete. This issue is made more complex by the lack of clear guidance from policymakers and the lack of consensus on how to balance considerations related to fairness, privacy, and representation.³

Incautious approaches to obtaining demographic data to measure fairness pose risks including misrepresenting people's identities, increased surveillance of already marginalized communities,

A clear-eyed assessment of the key concepts, synthesis of emerging demographic measurement approaches, and recommendations for practitioners and policymakers are urgently needed to enable responsible fairness measurement.

and dangerous misuse up to and including human rights abuses.⁴ Yet deferring too strongly to privacy concerns risks hindering anti-discrimination efforts,⁵ and the imperative to root out bias and discrimination in AI-powered systems has become firmly embedded in norms, policy proposals, and regulatory actions, leading some organizations to try to obtain data in order to enable what they understand to be responsible practice or compliance. In fact, researchers and some industry actors are already piloting and deploying various methods in this direction.⁶ Given the interest from advocates, policymakers, and industry in making progress on these issues, a clear-eyed assessment of the key concepts, synthesis of emerging demographic

measurement approaches, and recommendations for practitioners and policymakers that account for practical needs and constraints are urgently needed to enable responsible fairness measurement.

This report provides methodologies, guidance, and case studies for those undertaking fairness and equity assessments — from approaches that involve more direct access to data to ones that don't expand data collection.

- **Practitioners** in both public and private sector contexts will find a landscape of demographic measurement methods they can consider, examples of these approaches deployed in practice, recommendations around technical and organizational safeguards that should accompany this work, and call-outs throughout the report to help responsibly navigate these and similar approaches.

- **Public interest stakeholders** will find a summary of existing and emerging practices across the public and private sectors, learn about instances in which these approaches are already in use to inform continued advocacy, and understand implications of organizations using different methods for addressing bias and enforcing existing laws.
- **Policymakers and regulators** will find insights about the tradeoffs inherent in fairness measurement, insight about methods that may be useful for enforcement activities, and examples of practices that covered entities should explore or consider adopting, which can inform policy deliberations and support accountability efforts.

Section 2 highlights key decisions practitioners face when measuring fairness, including whether to look at impacts related to real people or to representations of people, and deciding which characteristics to measure. Section 3 describes different ways practitioners can uncover demographic patterns, with subsections on each approach including a description, examples of use across the public, private, and research sectors, and important considerations to keep in mind. Section 4 discusses approaches to responsibly handling the sensitive data and methods used in these measurements, again including an overview, examples, and important caveats. The report concludes with actionable recommendations for practitioners and policymakers to enable more of this important work while navigating its risks.

This report does not directly tackle the role third party researchers can and should play in helping to drive more awareness of harms communities may be facing, but discusses an array of examples, from organizations testing their own systems, to researchers probing systems with varying degrees of access, to enforcement bodies seeking to hold actors accountable for adverse outcomes; many researchers will find insights throughout that can help facilitate their important work.

Before proceeding, it is important to note that framing measurement as part of efforts to simply “debias” systems can lead practitioners and investigators to overlook systemic issues, and can create the

problematic perception that conditions exist where technology can be neutral — when in fact all technical artifacts involve value-based choices. Measurement must be more than a source of metrics to optimize, but rather an exercise to inform both specific and systemic interventions that can address root causes of inequity and prevent discrimination. Responsible demographic measurement for AI fairness work takes care and nuance, but this report offers both a map and a compass to practitioners looking to navigate the challenges. Ultimately, we invite practitioners to consider these complexities as an invitation toward thoughtful progress rather than an excuse for inaction.



02

Understanding and Defining Relevant Dimensions for Measurement

Determining the Relevant Lens of Analysis

When measuring bias in algorithmic systems, practitioners first need to determine if they are measuring discrimination affecting real **people** with lived experiences, identities, and rights, or biases related to **representations** — depictions or references to people or identities that are not necessarily tied to specific individuals. For example, a system used to assess creditworthiness directly assesses and confers or influences a decision impacting specific individuals, while a system analyzing and categorizing images might primarily rely on portrayals of items or people depicted in those images. Though these concepts are unquestionably related (for instance, when content analysis or generation systems end up as components in systems that result in meaningful decisions about real people) and certain methods may be applicable across contexts, the questions that arise and methods for measuring disparities related to each can differ, so any effort to conduct bias testing should start by reflecting on which lens is relevant.



People

When measuring biases in automated decision-making systems or in the provision of an AI-powered service, a practitioner is likely interested in directly measuring disparities affecting *people impacted by those decisions or people trying to use that service*. Such an analysis involves questions like:

- Are women being offered less favorable loan terms than men when applying for credit?
- Are job applicants of color being hired at similar rates as white applicants by a particular employer?
- Is a plagiarism detection tool inaccurately flagging students for whom English is a second language at disproportionate rates?
- Are health insurance claims being systematically denied more often for patients with disabilities compared to other patients?
- Is a facial recognition ID verification system leading more Black users to be denied access to the service?

Measurements to detect bias or discrimination in these sorts of cases generally demand some awareness of demographic information about the real people affected by those systems (as well as relevant structures of marginalization that communities who share those characteristics navigate), in order to disaggregate relevant metrics or outcomes and look for disparities across protected characteristics like race or gender. When actual demographics are unavailable, analyses about real people tend to fall back to other data points as explicit stand-ins for demographic features of the affected populations. For example, when analyzing racial disparities in student loan access, researchers who lack access to race data about loan applicants might use patterns of segregation across zip codes to come to conclusions about Black applicants compared to the general population.⁷ In other cases, the same features might be used for more general insights about populations — for example, looking at patterns of disparities between applicants from zip codes with majority-white and majority-Black populations but not drawing explicit conclusions about Black applicants. (*The relationship between **inferences** and **proxies*** call-out box in Section 3 further discusses this distinction).

Representations

Meanwhile, efforts to measure representational harms tend to rely primarily on indicators related to *content* or *representations* of people, rather than specific individuals. Analyses to detect these sorts of bias can involve questions like:

- Do internet image search results for “CEO” mostly return images of men?
- Does a translation system improperly assign gendered pronouns to genderless words in a way that reflects gender stereotypes?
- When prompted with certain nationalities or religions, does an image generator return stereotypical representations?
- When evaluating hypothetical job candidates, does a large language model treat synthetic candidate profiles differently if a demographic characteristic is explicitly mentioned in the prompt text?
- Is an image cropping function less accurate for photos featuring darker-skinned faces?
- Does a content moderation system more aggressively demote content about social issues that are of particular interest to marginalized communities?

While systems reflecting these representational biases can impact real people when deployed, the root causes can sometimes be traced to component biases related to content and not directly tied to individuals. This means measurements of disparities related to content can enable meaningful progress even if data about people is more difficult to wrangle. For instance, to measure for potential racial bias, resume scanning systems could be probed with fictional names common to different communities to detect disparities even before exposing the system to real applicants,⁸ or photos labeled with apparent characteristics could be used to measure disparate performance without collecting data about people’s racial identity.

Teasing apart these approaches can offer creative opportunities to spot patterns of bias, but it’s important to remember that though signals about content and **surrogate characteristics** may seem further removed from individuals, mishandling them could still negatively impact people.

Keep in Mind

When facing barriers to obtaining data about real people, practitioners can explore assessing disparities related to content or **surrogate characteristics** — that is, characteristics correlated with protected characteristics, but not necessarily linked to or intended to reveal sensitive details about specific individuals. This can raise fewer privacy concerns while still being relevant to systemic inequities faced by people who share those characteristics.

For example, an organization looking to understand potential racial disparities in a recommender system's content distribution could use the *zip code where content is produced* as a surrogate characteristic, since zip codes correlate with racial demographics due to histories of practices like redlining. Even though this analysis cannot directly answer whether the system imposes biases on Black users, findings would still be valuable to understand geographic disparities (which tend to align with racial disparities) and provide an indication that race-based bias likely exists even if difficult to directly measure.

Practitioners should avoid using content-level or surrogate characteristics to assign characteristics to individuals for measurement purposes. In the zip code example, the racial makeup of zip code populations should not then be used to assign racial categories to individuals associated with those zip codes, as this can obscure effects on underserved communities not in the majority in their zip code. Care should also be taken to ensure content-level or surrogate demographics are not inadvertently attached to users for purposes like ad targeting or consequential decisions.

See Section 3 for further discussion of proxies and surrogate characteristics, and Section 4 for more about approaches to handling data and methods to prevent misuse.

In some cases, practitioners may need to consider disparities related to opportunities allocated to *organizations or entities*, rather than individuals — for example, a government agency looking at potential biases in allocating small business loans, a content platform considering fairness in distribution of posts from various news outlets, or a company considering equity in their business-to-business fraud detection practices. For instance, researchers at Spotify described challenges in how to categorize “ensemble-based creators” like musical groups or podcast teams when considering fairness by gender. In such cases, practitioners face questions about whether to consider an organization’s “demographics” based on the characteristics of associated people (like owners or organizational leadership), or to categorize organizations based on intrinsic indicators like their topic, location, or size. While there may be justifications for either approach, practitioners should thoughtfully deliberate on which will be most feasible, accurate, and useful in identifying unfair outcomes. They need to decide if conferring people’s characteristics to the organization, or using the organization’s own attributes, is the better path for that particular analysis.

Selecting Demographic Dimensions and Segmentations

Once practitioners have identified the relevant lens of analysis (focused on real people or representations), they next face questions about which demographic dimensions and identities to consider in their measurement efforts. While sociodemographic identities can be complex and fluid,⁹ there are often historical and legal reasons to rely on defined categories. Many of the measurement approaches described later depend on having categorical demographic data, and in many cases, even continuous

Before initiating data collection or measurement — especially if a chosen approach deviates from norms — it is critical to involve external experts, intended system users, and impacted communities.

variables like age or income or nuanced identities like gender identity tend to be clustered into categories to facilitate comparison across a smaller set of groups. In other cases where theoretical justification for categorization is less strong, artificially grouping people can risk masking harms, particularly for people more similar to peers in other groups than to those in the one they have been assigned. Decisions about which predefined categories and subcategories to include, how granular they should be, and whether to allow people to relate to multiple subcategories (for example, people who identify with multiple racial identities) will inform and be shaped by feasible measurement methods. Accordingly, before initiating data collection or measurement — especially if a chosen approach deviates from norms — it is critical to involve external experts, intended system users, and impacted communities. These foundational conversations will profoundly affect those communities, so their involvement is essential.

First, organizations need to determine which demographic characteristics to measure. Options include:

- Only characteristics explicitly protected by law (e.g. race, religion, sex, national origin, age, disability, etc.);
- A subset of legally protected characteristics for which data is more readily available (race, national origin, gender, and age are often prioritized);
- Other socially salient characteristics beyond legal protections (such as language spoken or socioeconomic status); and
- Characteristics particularly relevant to the system in question or for which errors seems to be evident (e.g. accent for voice assistants, hairstyle for face detection).

Here too, input from external experts and impacted communities is crucial to understand what characteristics are relevant.

Once high-level demographic characteristics have been identified, practitioners must define how sub-categories will be partitioned. For example, if including race and ethnicity, they could:

- Rely on sub-categories defined by government sources like the Office of Management and Budget or Census Bureau (e.g. White, Black, American Indian, Asian, Pacific Islander for race; Hispanic/Latino or Not Hispanic or Latino for ethnicity);¹⁰
- Combine race and ethnicity into one overall race/ethnicity category, still using government-defined groups;¹¹
- Start with government categories but rename some groups, add categories (e.g. Middle Eastern¹²), or offer more granular sub-categories (e.g. Chinese, Filipino, Japanese, Korean and other Asian rather than one category for Asian/Pacific Islander);
- Work directly with communities to craft a set of characteristics that are both relevant to the measurement context and avoid representational and other harms.

In any of those cases, the approach chosen will depend on the context and input from impacted communities on which categorizations are most appropriate and useful for identifying potential discrimination. Some regulators explicitly narrow their focus; for instance, Colorado's Division of Insurance has opted to require insurance providers to include only the race/ethnicity subcategories of Hispanic, Black, Asian Pacific Islander (API), and White,¹³ while the Census Bureau has moved toward including more detailed race and ethnicity options.¹⁴ (While this example is pertinent to race and ethnicity in the US specifically, similar questions may arise for other demographic groups).

Though official demographic categorizations from sources like governments have faced criticism over the years for lacking inclusivity and not reflecting lived experiences, organizations relying on preexisting data or methods, as well as government agencies conducting equity analyses or enforcement may have less flexibility and need to rely on these existing taxonomies. Other entities find themselves navigating a more ambiguous landscape when it comes to demographic measurement. For them, engaging with external stakeholders and impacted communities becomes even more crucial: input from these groups is important to ensure that relevant efforts serve the needs of the communities facing discrimination that measurement is intended to detect.

Finally, practitioners need to determine whether and how to aggregate the demographic information they seek. For example:


- If using granular categories may not result in enough data in each for statistically significant analysis, they may need to combine — or drop — categories for analyses.
- For methods that result in people relating to multiple sub-categories or probabilistic predictions about characteristics rather than definite categories, they'll need to decide whether to simplify the data for basic analysis or retain the more complex signals.

A responsible approach balances statistical rigor with preserving important nuances. Contextual experts will have important insights into how these decisions might illuminate or mask important effects on impacted communities, and data scientists and statisticians can help ensure that the resulting analysis is sound. For instance, if data follows a statistically normal distribution within an aggregated category, using that level of aggregation might be helpful to inform policy or technical interventions. If data happens to be skewed within categories, though, then aggregation can be misleading and efforts to mitigate apparent disparities could fall short.

The following sections describe approaches for each lens of measurement in turn, as well as methods that span contexts. In particular, we distinguish between approaches to **measuring** demographic characteristics from questions of **handling** data reflecting those characteristics. While elements of the two are often conflated, approaches of each type can be mixed and matched, and by disentangling them practitioners and stakeholders can more clearly deliberate about potential paths forward.

Section 3 describes the measurement methods: approaches by which organizations may obtain, observe, access, impute, or otherwise understand demographic characteristics, approximations, or patterns. Descriptions of approaches and provided examples are meant to be illustrative and do not necessarily constitute blanket endorsement of that approach or its use. For each approach,

a discussion of caveats and limitations is provided. Section 4 highlights common and emerging approaches to handling this demographic data responsibly — the methodological, technical, and organizational guardrails practitioners should consider when integrating bias measurement. References to specific organizations do not endorse their practices, but rather reflect how much public information they have shared.*

 *

The author previously worked at Meta, including on efforts referenced throughout this report. This practical experience in the domain of AI fairness and demographic measurement informed the report's overall scope and analysis.



03

Approaches for Measuring Demographic Characteristics for Fairness Measurement

Practitioners aiming to measure demographic biases or disparities first need to identify potential approaches to access relevant data about the population or data in question, and choose one or more approaches to adopt.

While this phase is often framed as “collection”, we refer to this exercise more holistically as **measuring** demographic data since a variety of methods beyond rudimentary data collection exist that can facilitate disparity measurements related both to people and representations. We identify five prominent approaches to revealing disparities related to people: **collection, observation and inference, proxies and surrogate characteristics, auxiliary datasets, and cohort discovery**. Methods to measure disparities in representations, meanwhile, include **keywords and terms, and observation and labeling**. We also highlight approaches used across contacts: **synthetic data, exploratory analysis, and qualitative research** — which in particular is uniquely capable of illuminating lived experiences and contextual information to inform fairness assessments and identify potentially useful interventions



both within and beyond the four corners of an AI system. For each approach, the report describes the method, identifies some known uses by public and private entities, and discusses relevant considerations and implications in order to outline potential paths forward while examining the tradeoffs of each approach.

While each approach is discussed separately to support clarity, in practice, multiple methods can be used in conjunction. Such mixed-methods approaches that combine qualitative and quantitative insights can help surface important insights about an AI system's impact and its implications.¹⁵

Measuring Disparities Related to Real People

Collection

Directly asking or giving individuals within a population the opportunity to share their identity characteristics

When faced with the challenge of assessing bias without adequate data, a common instinct is to consider the simple **collection** of new data. This typically involves directly soliciting individuals or providing them with the opportunity to disclose their identity characteristics, which are subsequently compiled into a dataset.[†]

† “Collecting” can be used to refer to a broad variety of attempts by an organization to access data they don’t have; we use a narrower definition for the purposes of this paper to more clearly distinguish between practices that are often lumped together.

Collection typically involves directly soliciting individuals or providing them with the opportunity to disclose their identity characteristics, which are subsequently compiled into a dataset.

In some cases, organizations may choose or be mandated to attempt **comprehensive collection**, where every individual within a specified population (e.g., applicants, users, students) is given the opportunity to share specific demographic details.

This is usually via a structured form or survey, and the resulting data can then be used to report aggregate statistics to regulators, or to self-test relevant systems or decision procedures against quantitative nondiscrimination requirements.

Comprehensive demographic data collection of this nature is frequently driven by legal requirements.

For example, provisions of civil rights laws and associated regulations require US employers and mortgage lenders to give all applicants the chance to self-identify into predefined categories, albeit

with the understanding that responding to such questions is optional. Educational institutions are similarly mandated to gather demographic data to furnish aggregate statistics to the Department of Education, typically at the point of application or enrollment.¹⁶ While not necessarily required by law, some health providers also solicit demographic information from all of their patients to help address healthcare disparities.¹⁷

Other institutions opt instead for **partial collection**, asking a subset of the population of interest to self-provide demographic details in order to conduct statistical testing against the full population. Institutions may call for responses broadly until they receive a statistically significant and representative set of data, or they may construct a specific sample group or panel from which to collect demographic data. Partial collection appears to be common particularly among the growing group of tech companies who have taken action to begin addressing unfairness and discrimination.‡

For example, LinkedIn offers US-based users the opportunity to provide demographic information through a survey including racial

‡ This may be due to concern that attempts to collect sensitive characteristics from an entire population of users might be perceived negatively, particularly in light of preexisting concerns around corporate privacy practices and potential data misuse.

and ethnic identity, gender, transgender identity, sexual orientation, disability, age, caregiver status and military service.¹⁸ Users are not asked for this information upon signing up for an account, but can access the survey through their account settings and in-app prompts.¹⁹ The survey provides predefined categories from which users can select one or multiple options, but also allows write-in responses. YouTube similarly launched a voluntary demographic survey for creators, asking about gender, sexual orientation, race, and ethnicity. The company described the survey's goal as helping them “look closely at how content from different communities is treated in our search and discovery and monetization system” and “for possible patterns of hate, harassment, and discrimination that may affect some communities more than others”²⁰ — language that strongly suggests the survey results would be used to test AI models for bias. Other examples of technology companies and platforms that appear to have adopted partial collection methods as part of their approach to fairness measurement include Nextdoor,²¹ Meta,²² Instagram,²³ and Pymetrics.²⁴

Keep in Mind

Experts strongly recommend using self-identified demographic data when measuring and addressing disparities, since other approaches are less likely to reflect people's lived experiences and can lead to harmful errors and misclassification. But there are several important considerations for practitioners.

Privacy and transparency. Data collection raises a variety of important questions around privacy and transparency; a thorough discussion of these considerations can be found in Section 4.

Determining characteristics and data structure.

When collecting demographic data, organizations must first determine which characteristics to include, how to disaggregate them, and whether and how to handle free-response fields. These decisions are significant: determining who is counted, how categories are defined, and which

groups are excluded from data collection efforts has long been used as a way to exert power over populations, and so must be undertaken with great care.²⁵ Engaging directly with impacted communities is vital to avoid reifying socially defined categories,²⁶ misunderstanding the nature of certain characteristics,²⁷ and creating misrepresentation harms.²⁸ For example, in the context of disability, administrative or legal definitions can be significantly narrower than how people identify, and the use of those administrative or legal definitions can, as a result, lead to both failure to identify and meet the needs of disabled people, and erasure of their lived experience.

Microsoft appears to have recognize the importance of direct engagement, partnering with the Disability Data Initiative at Fordham University to engage directly with the disability community around expanding access to and use of demographic data to reduce disparities related to disability²⁹ — perhaps in response to the company’s research demonstrating that people with disabilities are open to sharing data if it will indeed help disability communities but are most comfortable doing so when disability-focused organizations were directly involved in the collection efforts.³⁰

Organizations must also navigate circumstances where characteristics resist binary or categorical definitions. For instance, the Office of the Chief Statistician of the United States released guidance for the collection of self-reported sexual orientation and gender identity (SOGI) data in federal statistical surveys, including example questions and resources for practitioners.³¹ Allowing free-response fields may help here, but decisions about how to preprocess those responses to enable analysis can be as significant as the initial definition of what categories to include.³² LinkedIn, for instance, describes in help text that “[w]hen you share demographic information on Self-ID through the free text boxes, LinkedIn may put the information you share into broader categories. For example, if you enter ‘Straight,’ we may put you into the category ‘Heterosexual,’ ‘Cis Woman’ into the ‘Female’ category, and ‘Korean’ to the ‘East Asian’

category.” Organizations should clearly communicate to users whether and how their self-provided identities may be reassigned into predetermined categories.

At the same time, in some cases regulators specifically allow the exclusion of groups smaller than a given threshold from measurements, meaning that aggregation can be an important tool to draw at least some attention to disparities. This is particularly important when considering whether to measure intersectional categories, which can quickly compound to represent very small cohorts. For example, New York City’s Local Law 144 requiring independent auditing of automated employment decision tools mandates auditors to measure the impact ratio of each tool (the selection rate or score of each demographic group against the group with the highest selection rate or score), and audits are expected to include measurements of intersectional groups (e.g. non-Latino males or Black females) — but the law also allows auditors to exclude categories representing less than 2% of the data,³³ a threshold that may well exceed the size of many intersectional groups, as well as particularly marginalized populations like Native and Indigenous groups. Clearly, decisions around aggregation and suppression of categories play a role in determining if certain identity groups will be visible in analyses at all, so practitioners should thoughtfully balance providing users agency over their identities and being able to effectively leverage provided information to detect and address disparities.

Coverage. Rates for voluntary demographic data collection tend to be low. Institutions and entire professional communities have made efforts to improve the rate and quality of data collection for equity purposes.³⁴ In the context of tech companies, LinkedIn published several blog posts encouraging users to self-identify,³⁵ while Instagram’s CEO recorded an explanation of the company’s race measurement efforts and called on users to participate.³⁶ Importantly, certain communities may decline to respond or misreport characteristics at disproportionate rates,³⁷ likely reflecting

the tension between desires for group visibility and privacy, especially for marginalized groups who have been harmed by formal data collection efforts in the past and worry that disclosing additional data will lead to similar negative outcomes.³⁸ These selection and nonresponse biases can contribute to misleading conclusions from quantitative analysis.³⁹ For example, misalignment around what constitutes a disability⁴⁰ and stigma around disclosure of disabilities may prevent the accurate collection of disability-related data.⁴¹ For race and ethnicity, low response rates are broadly understood to be detrimental to equity efforts, leading to a seemingly common practice of augmenting self-identification data with observation and inference methods (described in the next section).⁴²

Compensation. While efforts to improve coverage of demographic data in universal collection efforts often focus on enhancing trust through communication, training, and infrastructure improvements rather than compensation, some partial collection efforts have explored or offered compensation for individuals to provide demographic data. Meta, for instance, highlighted that participants in its Casual Conversations dataset (intended to help practitioners measure disparities in computer vision and audio analysis tasks) were paid for their contributions.⁴³ Individuals surveyed on the topic of disability data collection noted that compensation could be an added incentive for providing data⁴⁴ — but attempts to compensate people for providing “diverse” data can also lead to the collection of more data than needed, and contribute to exploitative or extractive dynamics.⁴⁵ For example, data annotation contractor Lion Bridge tried to offer Māori speakers \$45/hour to record audio in Māori, ostensibly to ensure functionality of an AI tool across languages. However, advocates for Māori data sovereignty see such efforts as a new frontier for colonization and would prefer to develop their own tools, or at minimum, engage in deep partnership with efforts to build for indigenous languages.⁴⁶

Observation and inference

Considering conspicuous traits or other relevant information to assign perceived demographic characteristics or otherwise predicting the likelihood of a person relating to a particular demographic group.

In cases where collection proves insufficient due to poor response rates or other limitations, organizations have looked to methods to **observe** or **infer** the demographic characteristics of the relevant population. Observation has traditionally involved considering conspicuous indicators such as visual or auditory characteristics to assign presumed or perceived categories to individuals. Inferring, meanwhile, can involve additional signals correlated with the demographic characteristics in question — via individual proxy characteristics, combinations of probabilities related to a handful of data points like surname and zip code, or more complex machine learning models — to predict the likelihood a person relates to a particular demographic characteristic.

Observation and inference are fairly common practices in some industries in the United States, and indeed are specifically envisioned by a variety of regulations, recognizing that people may opt against providing self-identified data but that disparity measurements are still legally required. For instance, the Equal Employment Opportunity Commission requires employers to submit workforce data through EEO-1 reporting; demographic data are drawn from employee records but rely on observer identification if employees opt against self-identifying. California's Racial and Identity Profiling Act (RIPA) requires law enforcement agencies to collect the observed age, race, gender, and apparent disabilities from people involved in any detention or search, including traffic stops. In other domains where collecting demographic data is explicitly prohibited, like consumer credit, observation and inference (sometimes called "imputation") is one of the only ways that regulated entities are able to assess their AI models for unlawful biases.

While observation and inference are often criticized for failing to recognize the complexity and inherent unobservability of identities, and for the faulty assumptions and inaccuracies inherent to common methods, they have nevertheless played a central role in enforcing civil rights laws in recent decades. Importantly, observed characteristics can be intrinsically relevant in certain circumstances, such as where discrimination may be triggered by perceptions of people's race or ethnicity rather than their lived identities.

The relationship between inference and proxies

The concepts of *inference* and *proxies* are often used interchangeably, but this report distinguishes them to enable a more nuanced analysis of responsible demographic measurement. *Inference* refers to explicit use of one or more data points to predict or assign demographic features to people. *Proxies*, meanwhile, refers to information that strongly correlates with a demographic feature, but is not then used to make explicit presumptions about group membership.

For example, the primary language or dialect someone speaks is likely a reasonable proxy for their national origin. Using language to predict or assign nationality to speakers of that language would fit this paper's definition of an inference. Using language to assess disparities faced by speakers of that language would fall into the *proxies and surrogate characteristics* approaches discussed in the next section. **This distinction is specific to demographic measurement for bias testing; it is not intended to justify incorporating proxies into model inputs simply because they are not being used to make explicit inferences about people.** Importantly, proxies for identity characteristics differ from actual identities, so practitioners should take care not to conflate them.

Credit and healthcare providers as well as government agencies commonly use Bayesian Improved Surname Geocoding (**BISG**), Bayesian Improved First Name Surname Geocoding (**BIFSG**) and related methods to infer race and ethnicity of populations using a combination of name and geography.⁴⁷ In fact, Colorado's Division of Insurance proposed regulations mandating that insurers use BIFSG to estimate race and ethnicity to detect unfairly discriminatory outcomes.⁴⁸ Financial institutions often estimate sex using first names and Social Security Administration statistics.⁴⁹ LinkedIn infers gender and age for all users on their platform, primarily for personalization and advertising but also for fairness and diversity efforts.⁵⁰ Airbnb worked with an external research partner to observe the perceived race of a subset of users in order to detect disparities in hosts' acceptance rate.⁵¹ A report describing Uber's civil rights and diversity efforts noted that the company's Marketplace Fairness team "has conducted limited collections or inferences of race or ethnicity data in order to facilitate the fairness testing of Uber's products," but did not elaborate on methods or scope.⁵² Some organizations may give people the option to opt in or out of such data processing activities to provide agency over data or comply with privacy and data protection laws where applicable.

Since observation or inference of sensitive characteristics can raise significant privacy concerns (for example, through the use of such methods for purposes like targeted advertising), organizations may look to adopt **privacy-enhanced observation or inference** approaches. While standard observation and inference approaches generally involve assigning demographic labels to individuals, and often storing those observations or predictions alongside individual identifiers for future use, some practitioners are experimenting with adding noise or aggregation to observation or inference procedures to address privacy concerns. Meta, for example, used the BISG method to impute race for US user populations, but incorporated several technical interventions to prevent individual-level predictions of race from being durably produced and to restrict the method to only outputting aggregate measurements.⁵³ A UK government study highlighted the importance of such methods, finding that methods drawing inferences at higher levels of aggregation could enable bias analysis without requiring service

providers to process individually-identifiable demographic data.⁵⁴ While some privacy-enhancing technologies like differential privacy can magnify fairness concerns,⁵⁵ make it harder to measure smaller groups who are more vulnerable to added statistical noise, or generate misleading findings due to the added randomness,⁵⁶ these methods have the potential to enable some progress toward bias measurement in a privacy-protective way where it would otherwise be impossible. They should continue to be actively considered and studied.

Notably, healthcare stakeholders appear to clearly differentiate between uses of inference for population-level insights, which is common,⁵⁷ and inferences about specific individuals, which is discouraged.⁵⁸ This is an important distinction with broad implications for fairness measurement, suggesting that population-level insights can be pursued even while creating durable individual-level data can be avoided.

Observation and inference are often combined with collection.⁵⁹ In some cases, methods may be adopted in parallel to develop a clearer picture of potential disparities, while in others approaches may be combined directly (which can complicate consequences for statistical analysis). Health plan providers use a mix of collection, imputation, and auxiliary datasets (discussed in a later section) to obtain race and ethnicity data. Meta, meanwhile, described adopting several complementary methods to measure race and ethnicity. The company relied on a paid panel of users who provided their race/ethnicity data to validate the BISG method's accuracy for analyzing the platform's data. The company also piloted a survey-based collection approach allowing a sample of users to self-identify, but articulated intent to augment the insights gained through collection with separate analysis relying on inference (using BISG) to account for anticipated low survey response rates that could lead to insufficient data to detect biases in key systems.⁶⁰

Keep in Mind

Observation and inference present risks both when methods are *in*accurate and when they are *too* accurate. While observation and inference may seem to be appealing alternatives or additions to collection approaches, practitioners should keep several caveats in mind and be extremely cautious before contemplating observing or inferring unobserved characteristics that could harmfully misrepresent or out people.

Accuracy. In recent years, some U.S. agencies that traditionally relied on visual observation for race and ethnicity have changed policies, now disallowing the use of observed race/ethnicity data due to inaccuracies affecting civil rights compliance.⁶¹ Research in the healthcare context found up to 66% error in observation-based approaches compared to self-reported data.⁶² On the other hand, some have noted this approach is particularly relevant when the kind of discrimination to be measured is based on perception.⁶³

Commonly used inference methods like BISG have also been criticized for accuracy gaps disproportionately affecting minority groups. Imperfect inference risks both minimizing disparities⁶⁴ and overestimating them,⁶⁵ introducing new biases.⁶⁶ However, research indicates that despite these limitations, inferential approaches are directionally informative and useful for making progress against disparities.⁶⁷ Researchers are exploring methods to account for statistical biases and uncertainty inherent in noisy imputation methods, which has tended to be overlooked in fairness analyses.⁶⁸

Researchers have iterated on these approaches to improve accuracy, creating BIFSG (incorporates first name),⁶⁹ fBISG (adjusting statistical imputation methods),⁷⁰ and simple machine learning.⁷¹ Alternative credit provider Zest, for instance, used a machine learning-based method to identify

patterns in name and location data to understand patterns related to race and ethnicity.⁷² But conclusions are mixed on whether including more features in predictive models provides sufficient marginal accuracy improvement to warrant their addition.⁷³ Given the many acute implications, practitioners are cautioned against pursuing new machine learning approaches to inference without clear, multistakeholder consensus.

Miscategorization and unobservability. While observation and inference have a long history in US civil rights enforcement particularly around race and ethnicity, practitioners should be extremely careful before using these methods to identify characteristics that lack such precedent or resist observation. Such approaches present greater harms of misrepresentation, miscategorization, and violating both dignity and privacy of people even in the pursuit of equity, so practitioners must grapple with whether and in what conditions they can be justified.⁷⁴ The UK Center for Data Equity and Innovation has suggested that inferring demographic data might be acceptable (with robust safeguards) when inferences allow more accurate bias identification than ground-truth demographic characteristics, when inferences are drawn in aggregate rather than at an individual level, or where no realistic alternative exists.⁷⁵ Even so, some characteristics like sexual orientation, some disabilities, and gender identity are invisible or fundamentally unobservable. Even if well-intentioned, attempts to guess them are highly problematic,⁷⁶ and can lead to nonconsensual disclosure of private information, imposition of stigma even if inferences are incorrect, or risks to physical safety where particular identities are considered unlawful. Even careful development of inference methods in close partnership with communities might appear to set a precedent for less scrupulous practitioners to justify more problematic methods.

These sorts of methods can also pose a risk of reinforcing stereotypes or stereotypical associations. Aggregate

inference approaches avoiding individual insights are a potentially promising path to explore, but not a cure for all risks. Deep, sustained engagement with communities must be integral to any exploration in this domain.

Consent, privacy and legal requirements. Despite established use for disparity measurement in certain contexts, online platforms' inference of sensitive characteristics for advertising and other purposes has raised significant privacy concerns in recent years, leading to both real and perceived barriers in the use of these methods to measure bias in digital contexts. Inferred characteristics may still be considered personal data under data protection regimes such as European or California privacy laws, with implications for what sort of **consent** may be required to process this data and the purposes for which this data may be used. Robust consent can protect against certain harmful data practices, and opt-in collection is viewed as a preferred approach to gathering demographic information, but offering people the option to give or withdraw permission to be included in population-level, inference-based measurements appears to be uncommon in most cases related to gender, race, and ethnicity. This is unsurprising, since different rates of opt-ins or opt-outs across communities can result in sampling bias and drive misleading conclusions about the existence or degree of disparity.

The EU has also taken action under the AI Act to strictly limit the development or use of systems that “categorize natural persons according to their political opinions, trade union membership, religious or philosophical belief or sexual life or sexual orientation” except in connection to criminal inquiries.⁷⁷ Some jurisdictions, such as France, reject the existence of racial categories as a matter of principle, effectively hindering the generation of related demographic statistics.⁷⁸

Operational details. If organizations use methods resulting in probability scores or distributions, they will need to

determine how to treat the outputs. For example, when using BISG for race measurement, a particular zip code-surname pair might result in data like 73% *Non-Hispanic White*, 22% *Non-Hispanic Black*, 1% *Hispanic*, 4% *Asian/Pacific Islander*. Some organizations, including the Consumer Financial Protection Bureau's Office of Research and Division of Supervision, Enforcement, and Fair Lending, rely directly on these raw probabilities to conduct statistical analysis. Others opt to classify individuals who exceed a preset threshold into the category with the highest probability before further analysis. (In the example here, the individual would be categorized as *Non-Hispanic White*).⁷⁹ Meta opted to rely on the latter approach, choosing a 50% threshold based on its own validation studies, precedent of use, and the method's interpretability relative to using raw probabilities.⁸⁰ In either case, inferred information should be stored separately from commonly used data, and should only be used for aggregate, statistical analysis. *See Section 4 for additional discussion of data handling safeguards.*

Proxies and Surrogate Characteristics

Signals or features that correlate with certain demographic characteristics (such as geographic location, occupation, or primary language) that can be used to detect patterns or disparities without directly making assumptions about individuals' demographics or making predictions based on these characteristics.

Proxies, or signals that correlate with a protected characteristic or other relevant demographic information, are sometimes used to explicitly infer demographic characteristics about an individual or

Proxies can be intrinsically useful to detect patterns that are likely to affect people with certain identity characteristics, without claiming that the proxy is analogous to the underlying characteristic.

population. But proxies can be intrinsically useful to detect *patterns that are likely to affect people with certain identity characteristics*, without claiming that the proxy is *analogous* to the underlying characteristic. For example, analysis might disaggregate a dataset or model performance by zip code-related characteristics (e.g. racial distribution of residents according to Census data) without assigning race/ethnicity to residents based on neighborhood statistics. Or, analysts might consider if users of accessibility features encounter more difficulty using an app/service, without presuming those users must be disabled. In some cases, disparities in experience or outcomes may even be primary interest, with protected characteristics themselves serving as proxies. For example, the Biden Administration's Climate and Economic Justice Screening Tool (CEJST) considers climate, environmental, and socioeconomic indicators to "identify communities that are shouldering a disproportionate share of environmental burdens and climate risks and that have suffered from underinvestment." The tool does not explicitly consider race, but the administration noted that communities of color are disproportionately affected by these impacts.⁸¹

While some may refer to such signals as proxies, **surrogate characteristics** is a helpful descriptor for these features, as they identify likely impact areas for unobserved demographics without assumptions about individuals. A surrogate characteristic analysis across zip codes would result in findings along the lines of, "the model appears to demonstrate bias against *loan applicants in majority-Black zip codes*." A similar analysis that reflected conclusions like "the model appears to demonstrate bias against *Black loan applicants*" would suggest that analysts used zip code to infer race based on geocoding, whether explicitly or implicitly.

Various features could theoretically serve as surrogate characteristics for bias analysis, such as geographic location at varying granularities, first and last names,⁸² languages spoken,

occupation, and type of device used to access an AI-powered service. Policymakers have already articulated an expectation that practitioners proactively search for features that are likely to be correlated with protected characteristics and to remove them from models powering automated decisions;[§] the same features may be useful as surrogate characteristics to measure model performance or outcomes for disparities when protected characteristics themselves are unavailable or insufficient. For example, Joy Buolamwini and Timnit Gebru used dermatological skin type classification to explore performance differences in facial analysis algorithms for darker and lighter skinned individuals, augmenting a dataset of global parliamentarians with phenotypic labels. They explained their choice of surrogate characteristics over observed race, describing that “subjects’ phenotypic features can vary widely within a racial or ethnic category” and that “racial and ethnic categories are not consistent across geographies: even within countries these categories change over time.”⁸³ They noted that since camera settings have long been calibrated in relation to skin tone, the characteristic would be particularly well-suited to use when studying image analysis algorithms.

Several social media platforms have alluded to using geocoding as a surrogate characteristic for bias analysis. Prior to its shift to new management and brand identity, Twitter researchers used coarse (county-level) location data and the racial characteristics of county populations to help measure racial bias in the platform’s algorithmic timeline. The researchers took care to present results as relating to population percentage rather than presumed user race, explaining that while insights from surrogate approaches differ from understanding individual-level disparities, they have been widely used to study racial disparities in domains like education and law enforcement.⁸⁴

§ Note that while such an exercise may also require demographic data in order to detect less obvious proxies, such an analysis may be more tractable using smaller datasets than an analysis of biases in a model itself or its outcomes. Third party stakeholders or auditors may also be able to provide guidance on which data points are likely to be proxies, and whether any of them may be appropriate to use as surrogate characteristics for fairness measurement.

**Surrogate characteristics
can help identify
likely impact areas for
unobserved demographics
without making
assumptions about
individuals.**

Meta has also alluded to the use of aggregate zip code data about racial patterns to conduct fairness analyses in the absence of individual-level demographic data.⁸⁵ The company appears to have used the CDC's Social Vulnerability Index, a surrogate signal based on census tract data and a variety of socioeconomic and health indicators, to measure and take action to advance equitable access to COVID-19 vaccines.⁸⁶ While this index incorporates features like neighborhood poverty levels, housing security, and lack of transportation (likely correlating with protected characteristics like race/ethnicity), the approach seems particularly suitable in this case since the index was crafted to relate specifically to health inequity and was used to detect and tackle potential drivers of health inequity. In other words, this method demonstrates strong *measurement validity*, or accurately capturing the concept it was intended to measure. In a civil rights enforcement capacity, the US government used the demographic characteristics of adjacent towns to investigate disparities in Meta's housing ad delivery.⁸⁷

Many industries and contexts appear to use surrogate characteristics, geographic and otherwise. For example, the financial company FICO has used zip codes to assess disparate impact of credit scoring for individuals living in "high-minority areas."⁸⁸ Researchers at the Mayo Clinic developed a method to combine addresses and publicly available housing data (e.g. home value and square footage) to understand disparities related to socioeconomic status.⁸⁹ Spotify appears to have considered disparities in artist popularity to detect and mitigate against "superstar economics" effects that disadvantage all but a few artists.⁹⁰ Journalists and researchers have also tested Amazon's and Google's voice assistants to see if they work differently for people with various accents.⁹¹

Keep in Mind

Like inferences and observations, proxies and surrogate characteristics can be easier to access, but their inherent imprecision means they too can minimize or obscure disparities across protected characteristics. For instance, a measurement comparing the experience of people living in majority Spanish-speaking zip codes with people in majority English-speaking zip codes could fail to pick up on biases that Latine people living in majority English-speaking zip codes may face.

The choice of proxy groups can also reinforce harmful stereotypes, leading to representational harms (meaning that engagement with and input from affected communities and contextual experts remains critically important). Still, some research has found that mitigating gaps for proxy groups that aren't directly related to protected characteristics can nevertheless benefit protected groups.⁹²

Using proxies and surrogate characteristics in a way that does not inadvertently slip into the realm of inference requires precise analysis. It can be all too easy to conduct an analysis using a surrogate characteristic like the racial characteristics of zip codes, but end up improperly drawing conclusions about the race of people living in those zip codes. To prevent this sort of slippage, analyses should specify the characteristic that was measured (e.g. zip codes with a majority Black population) and not revert to shorthand that suggests surrogate characteristics have been used to make inferences about individuals (e.g. labeling charts with data from majority Black zip codes as "Black").

Surrogate characteristics can be considered for their own merits rather than as stand-ins for protected groups. For example, measuring bias in generative AI detector tools using the TOEFL (Test of English as a Foreign Language) scores of students whose work is being evaluated by the tool would be immediately relevant to understanding

disparities that might affect non-native English speakers.⁹³ These scores could, but need not, be used as a stand-on for immigration status or citizenship. Relying on nonsensitive groupings within existing data can also help practitioners set up and test measurement pipelines for analysis and identify possible vectors of error that may not require demographic data to uncover, serving as both a measurement sandbox and independently providing useful insights about salient disparities. Using surrogate characteristics may alleviate concerns of practitioners or their legal teams that AI fairness work necessarily introduces potential legal liability — but may also be less useful for detecting or remediating violations of civil rights or equality laws, where the characteristics of interest are legally defined and insights from surrogates may be less actionable.

Auxiliary datasets

External sources of demographic data that can be combined with or compared against existing data.

Institutions and practitioners sometimes look to external sources of existing demographic data, or **auxiliary datasets**, to enable bias measurement. This can involve matching internal records about system performance or outcomes with an external source containing data about demographics, but not about the specific system being studied. Like partial collection, auxiliary datasets can be useful for validating inference methods to see if they are accurate enough for measuring disparities. In other cases, auxiliary datasets are obtained from partner institutions or publicly available data to enable analysis or fill gaps from collection-based methods. In some cases, a single entity may be made up of different bureaus, offices, or business units, each of which may not have complete access to

data held by the others; in these cases, even sharing data between different parts of the same organization can fall into this category.⁹⁴

Many institutions may hesitate to seek out auxiliary data due to legal or reputational risks of obtaining third-party data. Legal restrictions likewise limit federal agencies from sharing demographic information with each other.⁹⁵ Nevertheless, auxiliary datasets have played an important role in both the research and practice of detecting disparities, and the Biden Administration's Equitable Data Working Group has explored ways to lawfully compare across federal datasets, including Census data.

For instance, the Department of Treasury's Office of Tax Policy partnered with "other data producers to obtain microdata on race and ethnicity that can be used to validate the imputations when legally feasible."⁹⁶ Treasury also worked with researchers at Stanford to publish an analysis using an auxiliary dataset derived from publicly available voter registration records from North Carolina (which includes race) to validate the department's use of the BIFSG inference method by matching it with a subset of the tax data.⁹⁷ Researchers at Northeastern University, Upturn, and the Brookings Institution have similarly used North Carolina voter records to measure racial disparities in Meta's personalized ads delivery system.⁹⁸

Medicare health plans receive race and ethnicity data from the Centers for Medicare and Medicaid Services (CMS),⁹⁹ while other healthcare providers have reported receiving this data from employer records and state enrollment files.¹⁰⁰ Many healthcare stakeholders have urged more investment in infrastructure and guidance to support data sharing in order to help address limitations of other methods, like low response rates to collection-based approaches and inaccuracies of inference-based approaches.¹⁰¹

Keep in Mind

To join multiple datasets, each record needs a common identifier (or set of identifiers) to match records to one another. Personally identifiable information (PII) like emails, user IDs, names, social security numbers, and addresses are commonly used to facilitate this matching.¹⁰² Methods include linking records using PII and then removing PII before analysis, or encrypting the identifiers and matching records using the resulting hashed data.¹⁰³ When datasets don't share such a "key field" or the key field(s) are corrupted (e.g. due to missing and inaccurate data or changes in information over time), practitioners often turn to "fuzzy matching" — that is, matching records that are not identical using logical rules or machine learning to predict likely matches. Like any algorithm, an indirect matching method can lead to errors, and those errors can fall disproportionately on certain communities and contribute to misleading analysis.¹⁰⁴ For example, matching approaches relying on names could lead to more errors where data fields don't support multiple or hyphenated surnames or where a small set of names is common.¹⁰⁵ Organizations using auxiliary datasets should consider the effect of these biases in disparity analyses that rely on merged records.

Organizations facing a demographic data gap may be tempted to purchase data from third parties, especially if they commonly do so for other business purposes. While working with third party market research firms who fairly compensate study participants to provide data may reasonably balance considerations, **we recommend against companies or governments purchasing sensitive data from data brokers, who tend to unscrupulously collect and sell highly sensitive information about identifiable people, deploying substandard and sometimes unlawful privacy and security practices** — even if the intended use could seem to be positive.¹⁰⁶

Cohort discovery

Using manual or automatic pattern detection to look for clusters of people experiencing disproportionate errors or other negative outcomes.

Cohort discovery uses pattern detection approaches to look for clusters of individuals within a population who may be experiencing negative outcomes, without explicitly attempting to name or identify group characteristics that correlate with or have been captured by the clustering exercise. These approaches draw on the very nature of AI and machine learning that leads to concerns about fairness — that the technology can pick up on patterns that are difficult to observe — to measure AI systems for disparities that might be difficult to observe. The avoidance of descriptive labeling of subgroups differentiates this approach from machine learning-driven **inference and observation** methods described in previous sections, where a predictive model might be used to cluster individuals into predefined categories or assigned post-hoc descriptors. Indeed, clusters resulting from unsupervised learning-based approaches may or may not even correlate with demographic characteristics.¹⁰⁷

Cohort discovery cannot answer questions about whether an AI-powered system is unfair for specific, protected populations, but it can still be useful in identifying whether reasonably discrete populations may be experiencing disproportionate errors or other harmful disparities and support efforts to mitigate that harm. Practitioners could, for example, consider the cohort experiencing the worst outcomes or disparities and assess whether that level of performance or outcomes falls unacceptably short of the general population or best-off cohort, or aim to make iterative efforts to improve the circumstances for the worst-off cohort regardless of how that subpopulation stacks up against the average or best-off

groups.[¶]

Researchers in the healthcare context have piloted this sort of cohort-similarity approach, using machine learning to incorporate data points known to be associated with health care access and cluster patients into cohorts with non-descriptive code names.¹⁰⁸ While healthcare practitioners often have access to demographic data, the researchers hypothesized that demographics alone were insufficient to understand context-specific disparities and used cohort clustering to identify opportunities for supportive intervention and outreach to improve access to care. Researchers at Amazon evaluating disparities in the performance of the company's Alexa assistant similarly used automated cohort discovery to detect underperforming groups of speakers, clustering speakers not by demographic but by predicted model performance.¹⁰⁹

Researchers at Meta proposed a method for measuring fairness that relies on the concept of network homophily — that individuals who share sociodemographic characteristics are likely to be more closely connected in social networks. Their method avoids partitioning populations into discrete groups entirely, instead using mathematical measures of network distance as continuous (rather than categorical) indicators of similarity by which to conduct disparity analysis.¹¹⁰ Amazon researchers too reported promising results from similar “unsupervised” approaches, using statistical methods to compare model performance curves for facial recognition systems when demographic labels were unavailable.¹¹¹ While it is unclear whether either method has been deployed in practice, they may prove to be promising directions to explore.

¶ Traditional adverse impact tests might start by considering whether the selection rate for a particular protected group is less than 80% of the selection rate of the best-performing group; unsupervised clustering would still allow a measurement of whether the worst-off group fell below that rule of thumb — analysis would just not reveal with certainty whether the groups in question corresponded with protected characteristics. Practitioners could conceivably conduct a post-hoc or manual analysis using another method described in this paper to hypothesize or explore whether the least-performant cluster happened to overrepresent certain protected characteristics; this approach is discussed later in the “Exploratory Analysis” section.

Keep in Mind

While cohort discovery may help detect pockets of disadvantage that may or may not align with defined characteristics, clustering is vulnerable to assumptions data scientists may impose. The mathematical methods used to generate or identify clusters (e.g. *k*-means clustering) can lead to unstable findings and meaningless groups, particularly if clusters are not uniformly shaped. Clustering based on human review of data visualization can also generate misleading conclusions. Without the involvement of impacted communities or contextual experts, this method can be ripe for misinterpretation.

Unsupervised clustering approaches can help to identify cohorts experiencing disparities and may be useful to mitigate performance gaps by informing manual, iterative efforts to address issues experienced by underperforming cohorts or by facilitating more systematic oversampling of training data from those cohorts.¹¹² However, such methods will be intrinsically limited in their ability to reveal whether disparities intersect with protected characteristics in a way that would be useful for civil rights compliance or enforcement, or in ways that relate to social context and lived experiences.¹¹³

Measuring Disparities Related to Representations

The previous section described methods to conduct disparity measurements related to the demographics of actual individuals, but it may be possible to test many AI-driven systems for gaps that relate more directly to content, or indirect representations of individuals. For example, practitioners can measure **representations** of identity-related characteristics in datasets or generative system outputs, or probe language models for the use

of gendered assumptions without knowing or attempting to assign demographic characteristics to real people.

In certain cases, measuring disparities related to content or representations may be the most appropriate path to detecting bias — sometimes, it may even be a more direct approach. Importantly, when used judiciously, representation-based measurement can introduce opportunities to detect bias that may otherwise appear out of reach. For example, automated decision systems making determinations about content (e.g. content moderation or computer vision) may be better suited to measurements of disparities based on representation, whereas automated decision systems about people will ideally be measured in relation to characteristics of people. Nevertheless, where AI-driven systems that make decisions about people rely in part on content analysis, measurement of content-related disparities can be informative to questions of bias that people may experience even when demographic data about real people is unavailable to conduct the ideal set of measurements.

Keywords and terms

Manually or automatically constructing lists of words and topics that relate to demographic characteristics and using them to probe systems.

One common method to measure content-related biases involves manually or automatically constructing lists of **keywords and terms** that relate to demographic characteristics in question. Lists might include verbatim terms where demographics are explicitly named, or terms often associated with protected or sensitive attributes.

This approach is common in the testing of language models for biases and stereotypes. Sometimes, these datasets are constructed manually. For example, Adobe created a curated list of image generation prompts related to gender identity to test and remediate potential erasure of LGBTQIA+ identity and culture, particularly related to representation of drag queens, using the terms to test

for negative stereotypes or hateful depictions.¹¹⁴ Researchers have measured the extent to which mentions of demographic groups are associated with stereotypical professions¹¹⁵ or negative sentiment,¹¹⁶ generally using crowdworkers to evaluate whether model outputs reflect stereotypes. Another group of researchers constructed a set of 56 phrases describing people with disabilities to test whether natural language processing (NLP) models classified text differently or in a biased manner when disabilities were mentioned.¹¹⁷

Some have also turned to automated means to expand on an initial keyword or term list. Researchers from Meta built a dataset of more than 500 demographic descriptor terms across 13 demographic axes using algorithmic “nearest neighbor” analysis to identify terms likely to have a similar meaning. They then used a participatory process to solicit additional terms and feedback on automatically generated phrases. The research team also built a “demographic text perturber” to create variations of texts across characteristics like gender, race/ethnicity, and age in training and evaluation datasets,¹¹⁸ and used the resulting dataset to measure biases in several models including Open AI’s GPT-2 and Meta’s BlenderBot 2.0.

Keep in Mind

Measurements based on keywords and term lists will be fundamentally limited by the imagination of the people or automated processes generating those lists, and how thoroughly the process includes variations of terms or phrases that might also reveal undesirable disparities or bias. This includes the groups considered, the breadth of terms related to those groups, and the coverage of datasets across languages and cultural contexts.¹¹⁹ Robust participatory methods are critical to ensure that test datasets generated by this method are sufficiently inclusive and conducive to measurement validity. Without diverse input, keyword-based approaches risk overlooking important biases.



Figure 1. Demographic terms to test NLP (natural language processing) systems for bias from Meta AI.

Examples of identity-related terms representing multiple ethnicities, religions, races, sexual orientations, genders, and disabilities demonstrating how Meta researchers gathered and iterated on the terms to include in its bias measurement dataset.

Meta AI Blog, May 2022: <https://ai.meta.com/blog/measure-fairness-and-mitigate-ai-bias/>

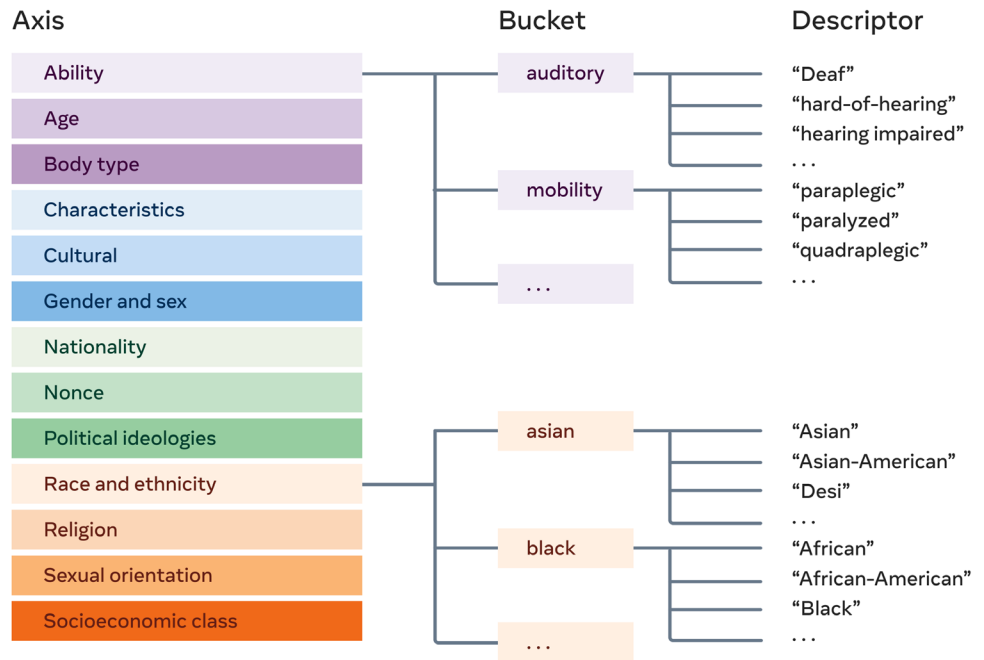
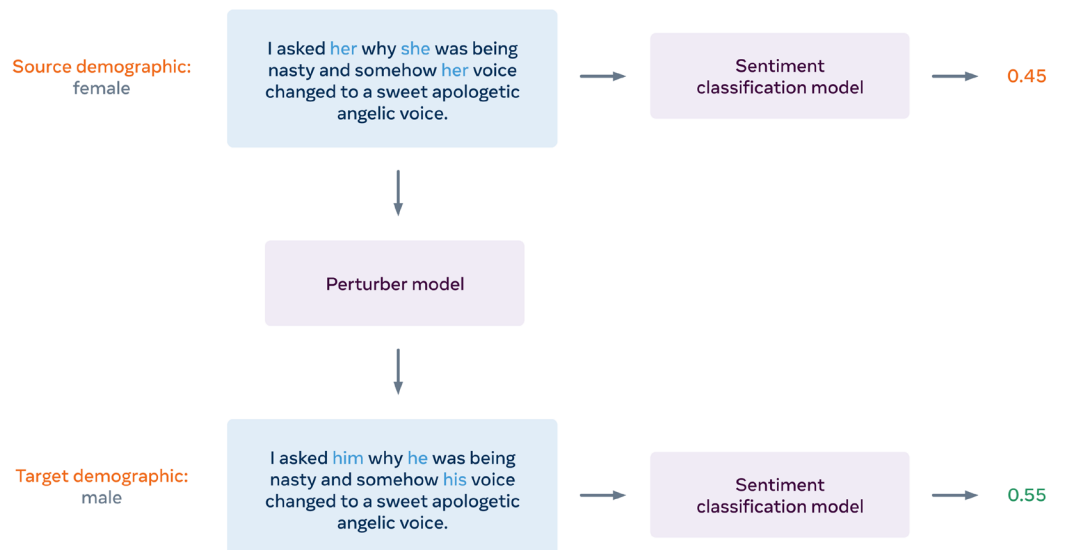


Figure 1 (continued). Demographic terms to test NLP (natural language processing) systems for bias from Meta AI.



Observation and labeling

Automatically or manually assigning labels of phenotypical or readily apparent traits to unidentified people represented in audiovisual or text content.

As with measurements of disparities affecting real people, *observation and labeling* appears to be a common practice for appending apparent demographic characteristics to content and representations. Practitioners frequently label phenotypic traits of unidentified people in photos and videos, as well as demographic-related stereotypes in generated content, to disaggregate measurements along those represented characteristics. In the case of content or representations, these labels tend to be attached to pieces of content like photos, videos, or text utterances, rather than assigning traits to specific individuals or based on personally identifiable information. While content items may have been created or shared by specific individuals, observation and labeling of content aims to only associate observed characteristics with the content itself, not individuals linked to that content.

Observation and labeling can occur manually, with data annotators reviewing and assigning perceived characteristics to each piece of content, or automatically, where labels are generated using an algorithm or model trained to predict whether a characteristic or set of features appears in a piece of content.¹²⁰

While much of such labeling takes place internally at companies developing AI technology, research published by industry labs illuminates what these practices look like. Google researchers, for example, asked annotators to assign labels related to perceived age range and gender presentation of unidentified people depicted in a computer vision dataset.¹²¹ Meta AI researchers used manual content annotation to construct the FACET dataset, labeling attributes of people depicted in images to facilitate the measurement of performance gaps of computer vision models across these traits.¹²² Journalists from Rest of World probed

image generation tool Midjourney for biases related to nationality by generating 3000 images with country-specific prompts and observing characteristics about the generated images including age, gender, skin tone, stereotypical attire, religious symbols, and the perceived socioeconomic characteristics of people representations in the generating images.¹²³

Pinterest deployed an automated model to identify body types and skin tones represented in images posted to its platform in order to increase representation in search results, initially using manually annotated images to train and evaluate machine learning models to identify these attributes. The platform shared that it now associates “all women’s fashion Pins with the prevalent body type present in them” and used the predicted characteristics to measure whether representation in search results improved following various technical interventions.¹²⁴ Sony similarly developed an automatic measure of apparent skin color, measuring skin tones across multiple dimensions** without classifying skin color into a predefined subset of categories to measure skin tone distribution in face-related datasets and discrepancies in generative model performance.¹²⁵ Another analysis went even further, removing hue and saturation from skin tone entirely and only focusing on the lightness of models’ skin tones in grayscale to evaluate representation in the cover of fashion magazines.¹²⁶

As companies build product experiences like virtual avatars, they may increasingly have access to structured, granular characteristics about those avatars (such as user-selected skin tone, facial hair, eyewear, use of hearing devices and other visual characteristics¹²⁷) — which may or may not reflect user’ actual appearance but nevertheless be useful dimensions along which to measure disparities in virtual spaces. We did not observe examples of companies disclosing the use of such avatar characteristics for fairness measurement, but would not be surprised if this new source of data has been considered, particularly to detect potential interpersonal discrimination in virtual spaces.¹²⁸

** Sony’s approach differs from some other skin tone approaches in that it expands on the light-to-dark spectrum by adding skin hue to its analysis, describing it as an important dimension to capture skin tones across communities.

Keep in Mind

Because the type of content typically at issue in fairness analyses of content-centric systems tends to be disconnected — or at least several steps removed from — identifiable people, our landscape analysis suggests that practitioners perceive observation and labeling of content to raise relatively fewer ethical and privacy concerns than observation and inference related to people. As such, it seems to be used often to measure biases in AI-driven systems.

However, similar concerns to those raised for observation and labeling of people still apply in the context of content or representations. Even if not assigned to individuals, assumptions about unobservable characteristics present ethical and representational harms, and even well-intentioned and thoughtfully conducted observation can set problematic precedent for other less scrupulous practitioners. For example, labeling images of people with perceived religion but constraining observations to only a few religions while ignoring others can contribute to pervasive invisibility of minority communities — and risks implying normative acceptability of systematically inferring people’s religion based on appearance. At minimum, practitioners should take care to clearly differentiate between perceived or implied characteristics and actual ones, and to avoid reverting to shorthand that could undermine such distinctions.¹²⁹

Even narrow efforts to observe reasonably objective phenotypical or descriptive characteristics present challenges: both manual and automated annotations suffer from reliability gaps,¹³⁰ the choice of categories can be ill-suited to the task (e.g. the use of dermatological skin types to measure computer vision systems),¹³¹ and inadvertent or improper linking of content labels back to individuals can present similar privacy concerns to the measurement methods discussed in the previous section.

Measuring Disparities Across Contexts

In addition to measurement methods that are specifically useful when considering disparities related to either people or content, several important approaches in the toolkit of fairness measurement have proven useful in analyses related to both contexts.

Synthetic data

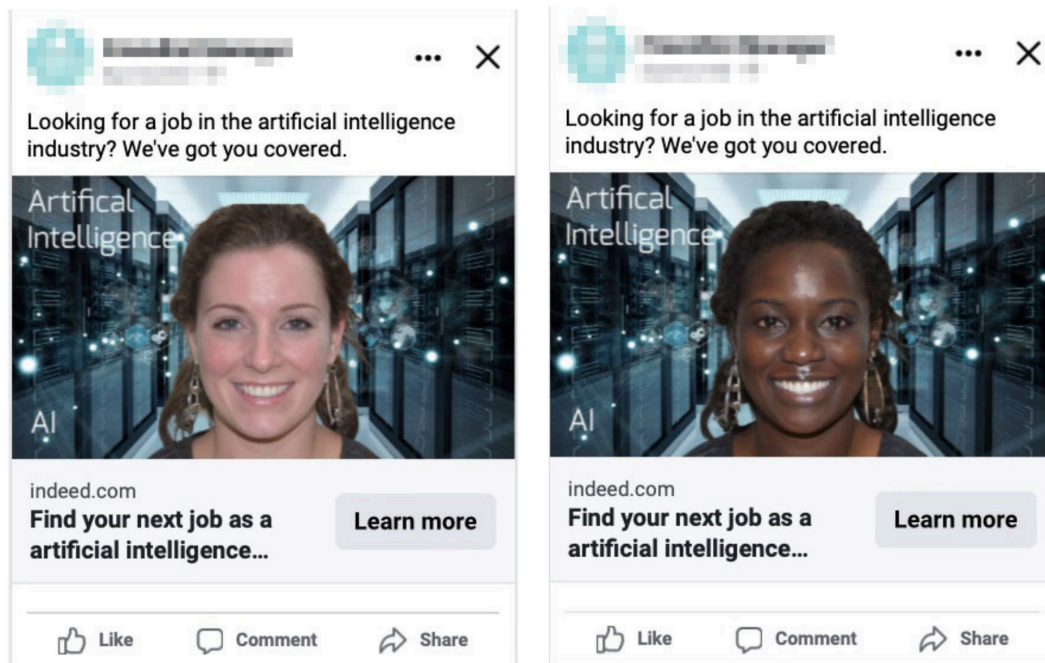
Artificially generated data that simulates the structure and distribution of real-world examples or populations.

Synthetic data is often discussed as a potential remedy for imbalanced and unrepresentative training data sets, but it could also help identify when such remedies are needed in the first place.

Synthetic data involves creating artificial data that is statistically similar to real-world data, without revealing personal or identifiable data.¹³² Synthetic data is often discussed as a potential remedy for imbalanced and unrepresentative training data sets.¹³³ But it could also help identify when such remedies are needed in the first

place. Approaches like synthetic datasets labeled with relevant demographic information can support the measurement of AI systems to find potential biases,¹³⁴ while methods like simulation studies can help practitioners explore counterfactual scenarios to test whether protected characteristics may have improperly influenced an AI system's outcomes.¹³⁵

A helpful analogy is civil rights testing and correspondence testing studies — when advocates, researchers, and certain regulators send fictitious applicants to probe for unlawfully discriminatory decisions about housing, government services, employment, and other life opportunities.¹³⁶ While in traditional testing, testers tend to use their own identities but simulate intent to seek the opportunity in question, researchers have also leaned



↑
Figure 2. Synthetically generated job ads.

Two examples of job ads researcher ran using synthetically generated faces to control for demographic representation.

Measurement and Analysis of Implied Identity in Ad Delivery Optimization, Proceedings of the 22nd ACM Internet Measurement Conference (IMC '22): <https://dl.acm.org/doi/pdf/10.1145/3517745.3561450>

on simulated application materials that control for qualifications but modify demographic characteristics to measure response rates. For example, well-known studies have involved generating and submitting fictitious resumes to job listings and measuring callback rates by perceived race.¹³⁷ AI researchers have turned to similar methods and synthetic data to probe algorithmic systems for bias.¹³⁸

Synthetic data has also been used to test for content- and representation-related biases. Here, instead of generating data reflecting realistic individuals that mirrors true statistical distributions within a population, synthetic representations and content are used to probe how systems analyze or generate content based on prompts. For example, researchers used synthetically generated ad images to test whether Facebook's ad delivery system distributed ads in a potentially biased way based on differences in ad creative — in other words, images, video and text used in the ad. Synthetic content allowed the researchers to hold all factors of the ad creative constant except for the demographic features of faces represented in the ad images, and then to measure the audiences each ad was exposed to.¹³⁹

In the realm of testing generative AI systems, researchers at Anthropic used language models to automatically generate an array of prompts related to decision-making situations to determine if Claude 2.0 gave discriminatory responses when different demographic characteristics were explicitly mentioned.¹⁴⁰ The researchers were able to use the same dataset to test the effectiveness of approaches to mitigate against such behavior. Microsoft researchers similarly produced a dataset of AI-generated texts (supplemented by human annotation) to help measure fairness harms related to gender and sexuality that language models might cause, and used this synthetic dataset to test GPT-2 for such biases.¹⁴¹

Keep in Mind

Synthetic data can offer a tempting option to overcome barriers in existing datasets, but it is more complex than it appears. Generating and manipulating artificial data can require statistical expertise to ensure it closely mirrors real-world patterns — particularly when synthetic data itself is generated by statistical models. Synthetic data is also not immune from its own biases, such as ignoring or underrepresenting some communities and overrepresenting others.¹⁴² Addressing this can require active efforts to reduce biases at the point of data generation like reweighting and oversampling — which itself may require access to demographic data of an underlying population, particularly when used to generate synthetic records about people rather than content. (If an organization has opted to use another demographic measurement approach, synthetic data based on insights those approaches generate could be considered as part of efforts to protect privacy and prevent misuse of individually identifiable data.) Some researchers have warned, though, that using generated data for fairness testing compounds measurement uncertainty, because tests end up blending the effects of both the model that has generated synthetic data and the model(s) being tested.¹⁴³

Incautious reliance on synthetic data also creates risks of applying narrow technical interventions to problems with sociotechnical dimensions, for example by creating datasets that overlook real-world constraints that may be driving disparities or by diverting attention from valuable qualitative insights.

When it comes to content-oriented measurement, it's important to remember that just because generated data may be synthetic, it can still rely on stereotypical assumptions and contribute to representation harms, via choices like the selection of categories to be represented and how those categories are defined. The process of labeling synthetically generated data also raises similar concerns as labeling organic data. Researchers have attempted to be sensitive to these considerations while still making progress in detecting bias; for instance, the ad research project described in this section used an automated method to estimate gender, race, and age of representations in the images they generated, but acknowledged in-line the limits of binary gender categorization, the socially constructed nature of race, and the fact that race cannot be observed from images, while other researchers emphasized the importance of diverse annotators to contribute labels characterizing the generated data.¹⁴⁴

Exploratory analysis

Reviewing a system's data, artifacts, and outcomes to reason about how the design, behavior, or other characteristics of the system might lead to negative impact for certain communities.

Even when structured data is not available for disparity measurements, there are still ways to look for unfair differences.

Exploratory analysis of data can reveal clusters of errors that, based on qualitative reflection on the type or apparent cause of those errors, clearly impact a particular community at disproportionate rates.¹⁴⁵ For example, data analysis as part of a human rights impact assessment of Meta's policies around violence in Jerusalem in 2021 revealed that posts about the Al Aqsa mosque on Instagram were incorrectly determined to be related to a similarly named organization on the company's list of "violent or dangerous organizations." The assessor's analysis suggested this error likely contributed to "an adverse human rights impact... on the rights of Palestinian users to freedom of expression."¹⁴⁶ In the context of employment, exploratory analysis might include the use of explainability tools like salience mapping to identity elements of a resume that a screening model relied on to inform its prediction to determine if a model inadvertently ingested irrelevant information like candidate names or gender-coded sports.¹⁴⁷

Creating more interpretable models can also reveal problematic drivers of errors. By manually examining factors that seem to be important to a model's outcomes, practitioners can reflect on whether those drivers might disproportionately affect certain populations. Explainability requirements can help discourage reliance on discriminatory factors and spot errors.¹⁴⁸ And testing model robustness — exploring under what conditions a model fails — can also highlight relevant failure modes. For instance, some researchers have explored approaches to systematically optimize models to prioritize improving worst-case scenarios,¹⁴⁹ and others have tested efforts to oversample data points that seem to be resulting most often in errors, regardless of whether errors were associated with demographic characteristics.

Keep in Mind

Since bias can be understood as errors that disproportionately affect certain communities, focusing on errors that seem likely to affect underserved groups can directly address issues leading to unfair outcomes. This approach can directly inform improvements or interventions without relying on demographics.¹⁵⁰ Analyzing errors while avoiding sensitive characteristics altogether may help organizations more easily incorporate systematic measurement efforts across an organization without fearing misuse of demographic data. At the same time, the connection between patterns of errors and protected characteristics may not always be clear, which could pose challenges for efforts grounded in civil rights laws.

Qualitative research

Direct engagement with people using and affected by systems to capture more nuanced insights about people's lived experience that quantitative methods may overlook.

Qualitative methods must not be overlooked in the toolkit for understanding demographic patterns. As the National AI Advisory Committee affirmed, "AI systems are sociotechnical systems and should be studied as such," which require consideration of human, social, and cultural contexts that surround technical artifacts.¹⁵¹

As such, direct engagement with marginalized people and communities is critically important to identify potential harms and unfair impacts that they face or anticipate facing from AI systems.

Qualitative research methods like ethnography and user studies have long complemented quantitative measurement, and differ from

exploratory analysis in their depth and structure, and in many cases, a fundamentally human-centered approach resulting in what some call “thick data.”¹⁵² While qualitative methods can’t deliver statistical evidence of systematic biases or precisely quantify whether an intervention has sufficiently closed gaps, they provide critical context and remain vital tools of AI practitioners to investigate and prevent bias, especially when quantitative data is limited.¹⁵³ The White House Blueprint for an AI Bill of Rights clearly notes that proactive equity assessments of algorithmic systems should include both qualitative and quantitative evaluations in order to identify potential discriminatory behavior or effects.¹⁵⁴

Studying AI-driven systems holistically — not just the models that power them — can reveal flawed assumptions, policies, and processes that can lead to disparate effects.¹⁵⁵ Advocates recommend examining the people and institutions who build and deploy these systems through a lens of power structures and institutionalized oppression.¹⁵⁶ For example, scholars studied Facebook’s “real name” policy by reviewing public documents, policy documentation, and statements from company leaders to show how it disproportionately excluded users with non-normative identities.¹⁵⁷ Statistical evidence may not be needed to recognize how surveillance technology is likely to have an outsized impact on Black, brown, and poor communities, due to its carceral effects.¹⁵⁸ An ethnographic study on predictive policing in Delhi overcame barriers to quantitative measurement and quickly uncovered that crimes were over-reported in so-called ghettos and slums — clearly leading to a criminalization of poverty — and that the intended law enforcement users of the technology did not even understand the system.¹⁵⁹ The US Department of State worked with the National Center for Health Statistics to conduct cognitive interviews about how passport applicants would prefer to indicate their gender on legal documents, leading to the adoption of the “X” gender marker on a number of federal forms.¹⁶⁰

Many tech companies use ethnographic methods to evaluate aspects of their products, including those driven by AI; in industry settings, this is commonly referred to as user research.¹⁶¹ Some companies have established “product equity” teams to focus

more closely on marginalized populations' needs, at times even co-designing products or policies with those communities.¹⁶² For example, Adobe has described its intent to develop “longitudinal and in-depth understanding of prioritized social identities” to uncover gaps in the company’s understanding.¹⁶³ Google Pixel’s accessibility group conducted qualitative testing with blind colleagues to test usability of a camera tool for blind and low-vision people, and engaged in hands-on testing together with creators to identify potential issues with lighting and contrast quality for darker skin tones in different photography modes.¹⁶⁴ Researchers from DeepMind cited the importance of safe spaces for marginalized identities, such as the Queer in AI workshops at major machine learning conferences, as opportunities to discuss impacts of AI systems for demographics not readily observed, such as sexuality, religion, disability, and class.¹⁶⁵

Keep in Mind

Many companies already employ user researchers, who already engage in qualitative research and are both poised for and well-suited to apply ethnographic approaches to the challenge of identifying unfairness. In order to ensure qualitative research includes perspectives from communities most likely to experience negative impacts, organizations often must actively recruit diverse study participants — which can be difficult without demographic data to help identify and reach potential participants. Even if such data exists, targeting advertising or opportunities using that data is often disallowed. Faced with such constraints, some organizations work with outside firms who specialize in inclusive user research who have existing relationships and trust with underrepresented communities. For broader research initiatives, the lack of demographic data can impede stratified sampling to understand the validity of research results findings across communities.¹⁶⁶ Finally, qualitative studies can be resource-intensive and challenging to track over time. But given the limitations of rigid demographic categories via surveys or indirect measurement

methods discussed throughout this report, qualitative research is hugely important to inform and augment efforts to understand disparities across communities, providing important context both on its own and as a complement to other measurement efforts.



04

Approaches for Handling Demographic Characteristics for Fairness Measurement

Distinct from the methods used to obtain information about demographic characteristics of people, populations, or content are questions about how that data and measurement techniques are set up and governed, both technically and procedurally. We call this set of practice the **handling** of demographic characteristics, which can include an array of privacy and security interventions, infrastructure and access controls, and policies or procedures that govern how data and methods can be used.

Demographic data can be highly sensitive; its misuse and leakage over the years have led to instances of widespread discrimination, exploitation, and violence,¹⁶⁷ which makes stakes high for the responsible and secure handling of this data. While discussion about data collection and measurement methods have matured in recent years, expectations around the proper handling of this sort of data remains notably unclear. Even advocates and practitioners who recognize the importance of demographic data to advance equity have noted the need for clearer guidance on allowable uses of data and confidentiality and integrity protections in order to build trust that sensitive data won't be improperly accessed or misused.¹⁶⁸

Where it exists, privacy and cybersecurity legislation provides some insight into these expectations, but often fails to specify the set of practices that would appropriately balance the need for fairness measurement with data minimization and protection.¹⁶⁹ The lack of comprehensive federal privacy legislation in the US or stronger civil rights guidance on the topic further complicates the situation.

The following sections outline some of the key technical and institutional tools organizations can employ when handling demographic data and methods. While not exhaustive, the practices described are representative of those that have been adopted or proposed in the context of measuring demographics.

Data and infrastructure controls

Pseudonymization

In light of concerns about revealing or misusing sensitive personal data, some organizations have turned to **pseudonymization** to obscure the identities of individuals associated with demographic information. Pseudonymization is a de-identification technique that replaces personal identifiers with placeholder information or otherwise breaks the link between identifying data and other data about an individual. Unlike anonymization, pseudonymization can be technically reversible and so still presents reidentification risk, but nevertheless makes it harder to attribute data to specific people.

In the context of demographic measurement, pseudonymization has been used to obscure both individual and group identifiers. For example, Airbnb's antidiscrimination team used pseudonymization, among other privacy interventions, in its efforts to measure acceptance rate disparities on its platforms, replacing internal identifiers with random identifiers prior to conducting analysis.¹⁷⁰ Similarly, when Meta launched a race measurement survey on

Instagram, they replaced internal user IDs with random IDs before having a survey provider solicit responses. The table linking user IDs and random IDs was stored separately from survey responses, preventing the survey provider from connecting responses to individuals.

Both Meta and Airbnb also describe efforts to obfuscate group identifiers. Airbnb replaced perceived race labels with stable, masked identifiers (e.g. group A, group B, group C). Meta encrypted race and ethnicity group labels in two ways: first, when using BISG, they “replac[ed] the sensitive feature name with a temporary random string to avoid making or storing explicit user-level inference of race.” Second, the survey encoded group names into 23-bit vectors before transmitting them.¹⁷¹

Infrastructure controls

Data architecture choices meaningfully impact how demographic data can and cannot be used. Organizations tend to build technical barriers to prevent sensitive data from leaking into systems where it is not permitted. The White House Blueprint for an AI Bill of Rights clearly states this expectation for such **infrastructural controls**, emphasizing that: “[d]emographic data collected for disparity assessment should be separated from data used for the automated system.”¹⁷² We observe three general approaches to data architecture choices when it comes to handling demographic data: *internal infrastructure controls*, *third parties and intermediaries*, and *federated architectures*.

Internal infrastructure controls. Organizations can build specific data stores, pipelines, or tools that limit how data and measurement methods can be used, and to prevent data from being disclosed to analysis. This can involve:

- Role-based access controls, where certain data tables are only visible or usable by certain job functions;¹⁷³
- Purpose-based access and use controls, which involve hard-coded limitations on how certain data can be used;¹⁷⁴

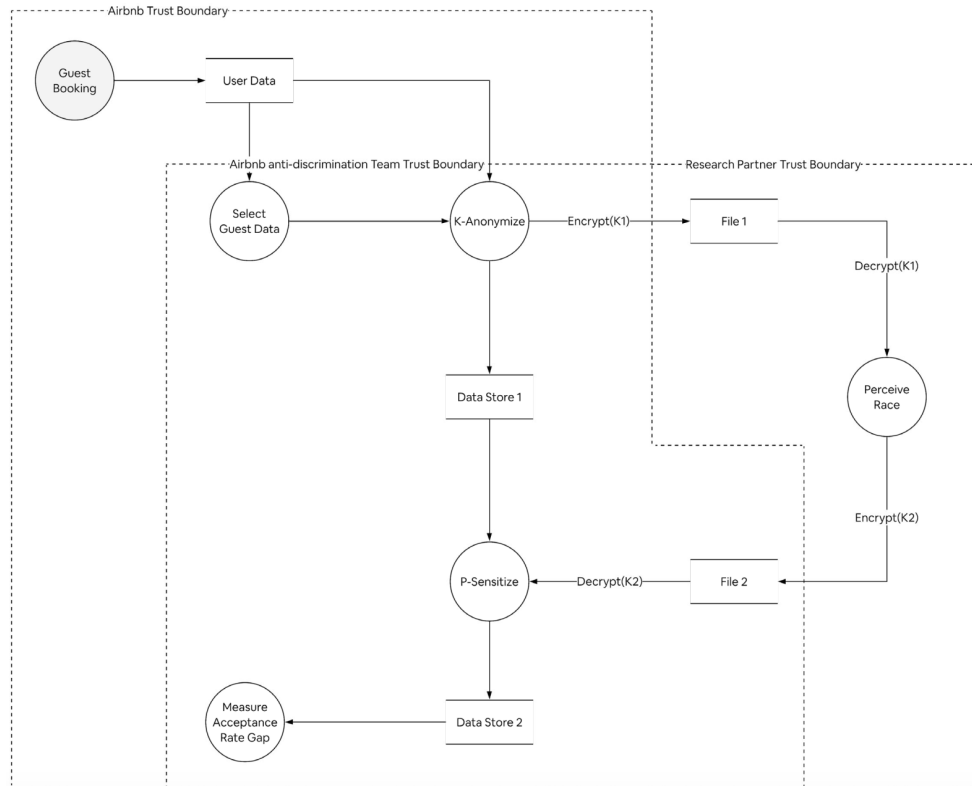


Figure 3. Airbnb Project Lighthouse simplified data flow diagram.

A schematic describing how demographic data flows within Airbnb and its research partner, including privacy protections and access controls.

Measuring discrepancies in Airbnb guest acceptance rates using anonymized demographic data, Airbnb, June 2020: <https://news.airbnb.com/wp-content/uploads/sites/4/2020/06/Project-Lighthouse-Airbnb-2020-06-12.pdf>

- Specific tooling that aggregates individual data into queries and statistics outputs without giving analysts access to individual-level data, by conducting predictions of demographics on the fly and immediately aggregating them without storing individual predictions.

For instance, Airbnb described using “trust boundaries,” which appear to be a form of role-based access controls. The company described them as “boundaries where the level of trust changes so that, for an actor to have access to data in datastores and access to run processes within a trust boundary, they would need to be authorized appropriately,” enforced by organizational firewalls and asymmetric encryption.¹⁷⁵

LinkedIn’s AI Fairness team describes building machine learning pipelines to enable model evaluation, without model owners needing direct access to demographic information.¹⁷⁶ Meta used a similar approach, implementing BISG through a contained tool that only provides analysts with access to aggregated measurement results rather than individual inferences.¹⁷⁷

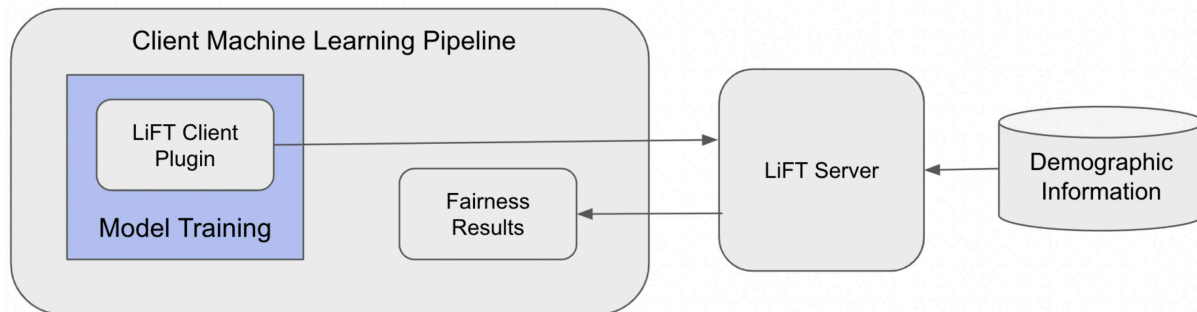


Figure 4. Fairness evaluation architecture at LinkedIn.

A schematic describing technical architecture that allows model owners to evaluate fairness metrics for their models without needing direct access to demographic information. (LiFT refers to the LinkedIn Fairness Toolkit, a system that houses measurement algorithms).

Disentangling and Operationalizing AI Fairness at LinkedIn, ACM Conference on Fairness, Accountability, and Transparency (2023): <https://doi.org/10.1145/3593013.3594075>

We did not find organizations publicly professing to rely on purpose-based limitations to safeguard demographic data for fairness measurement. However, technical approaches to comply with data protection regulations that require purpose limitation can involve this type of architecture.¹⁷⁸

Experts in the context of fair lending have warned that complete internal separation is not always feasible, especially for smaller organizations that may lack the resources and expertise to staff separate modeling and compliance teams. Data architecture choices affect how easy or difficult it is for teams to proactively search for less discriminatory algorithms, since locking data down too aggressively can significantly hinder teams from integrating proactive fairness measurements into their development process.¹⁷⁹

Third parties and intermediaries. AI developers who want to avoid holding demographic data or hosting measurement methods internally to prevent misuse or scrutiny may explore housing data or methods with third parties.¹⁸⁰ Potential data intermediaries include civil society organizations, law firms, market research or survey providers, data custodians, universities, and government-designated entities. These entities may already have demographic data about a population or may be newly tasked with obtaining it. In either case, methods to link that data with the AI developer's data about model performance are established, usually by creating a common identifier such as a name, email, numerical identifier, or a combination of data points. The data custodian can then be given access to models or relevant data from the developer to conduct

tests either directly or via API, or can implement more complex data sharing protocols like multiparty computation with the AI developer allowing data to be compared without data or model artifacts being revealed. The protectiveness of such arrangements depends heavily on the security and data practices of the third party (e.g. prohibiting secondary use), and requires both due diligence and robust contractual terms to mitigate risks of sharing sensitive data with third party entities.

Research from the UK's Center for Data Ethics and Innovation found that 65% of survey respondents would be comfortable sharing demographic data with third-party entities in order to facilitate fairness measurement, trusting consumer rights organizations most and technology companies least to serve as data intermediaries.¹⁸¹ In a hypothetical scenario proposed by researchers, an insurer using machine learning to set premiums could work with a consumer rights group. The insurer would redirect users to the consumer rights group to provide demographic information linked with a unique identifier after soliciting standard insurance application materials. The consumer group could then access the insurance model to conduct disparate impact testing or identify other potential bias sources and relay findings to the insurer.¹⁸²

Meta piloted a variation on this protocol, working with a third party survey provider to solicit race and ethnicity data alongside a pseudonymized identifier. Survey responses were then cryptographically split and shared with several data custodians, allowing the company to conduct fairness measurements without accessing raw responses.¹⁸³

Federated architectures. As more people obtained personal devices, researchers developed methods to conduct distributed analyses across data stored on local devices.¹⁸⁴ Federated statistics (or federated analytics) “enables organizations to access and use data from multiple, discrete devices without the need to collect and store this data in a centralized database.” For demographic data, such approaches can help avoid creating centralized databases that may contain sensitive information and impose a strong

infrastructural constraint against accessing or misusing individual-level data.¹⁸⁵ Apple has piloted a version of federated statistics to spot biases in the identification verification steps of user onboarding for its IDs in Wallet product, combining distributed infrastructure with differential privacy (discussed in a following section) to prevent re-identification of the demographic data used to conduct algorithmic fairness tests. Apple has asserted that “no personally identifiable information is collected, stored, or used by Apple or the state issuing authority as part of this process,”¹⁸⁶ distinguishing this approach from most others discussed in this report.

Encryption

Encryption plays a crucial role in safeguarding sensitive data, both in transit and at rest, by scrambling data so it can't be easily deciphered without a mathematical key.¹⁸⁷ At a minimum, organizations should deploy some form of encryption to safeguard personal data, and sensitive demographic data demands even more care. Many organizations who have publicly described demographic measurement activities reference the use of encryption in some manner. For instance, in a help center article about its platform demographic survey, Nextdoor stated that survey responses would be stored with “bank-grade encryption.” Airbnb described its use of asymmetric encryption “to ensure that the internally identifiable data prepared exclusively for and sent to the research partner cannot subsequently be retrieved by Airbnb and re-linked to the [anonymized] data.”¹⁸⁸

Cryptographic techniques for obscuring data are also essential components of some privacy-enhancing technologies discussed in the following section, such as secure multiparty computation and homomorphic encryption. These techniques enable analysts to perform analysis and computations on a dataset without having access to the raw data.¹⁸⁹

Retention and ephemerality

In line with the concept of data minimization, data generation and **retention** approaches play an important role in balancing privacy and fairness measurement needs. Organizations should only collect and retain data that is necessary to accomplish a specific purpose and avoid keeping data longer than needed to fulfill that purpose. However, longer retention periods may be important to facilitate deeper analysis or monitoring over time while avoiding the need to repeat data collection or collect data from more people.¹⁹⁰ If policymakers expect practitioners to conduct measurements on a regular basis to account for model drift or re-test models when they undergo changes, for instance, limited retention periods could lead organizations to need to collect sensitive data over and over again, potentially negating the protections that shorter retention periods theoretically provide. Fundamentally, retention should align with user expectations and policy imperatives, and be communicated clearly. Further, retention limits should be paired with disposal methods that ensure data is permanently unreadable and unrecoverable.

Retention schemes can be technically hard-coded, set out by policy, or contractually defined with external organizations (such as data custodians). The Partnership on AI has recommended limiting the retention of demographic data for fairness measurement to 90 days.¹⁹¹ Airbnb and Meta described adopting a 30-day retention period for individual-level demographic data collected or observed as part of their respective race measurement efforts.¹⁹² Some organizations, including LinkedIn and Nextdoor, also allow individuals to modify or delete demographic data that they have provided via surveys.¹⁹³

Practitioners have also highlighted approaches that avoid durably attaching sensitive characteristics to individuals — or that avoid generating data altogether — while still enabling statistical analysis. Organizations may add friction to re-identification by maintaining a

pseudonymized table of quasi-identifiers mapped to demographic characteristics and a separate table of user IDs mapped to quasi-identifiers, only joining those tables temporarily within bundled queries that result in aggregate outputs.

Some have gone deeper in exploring ephemeral measurement. For example, the race estimation method BISG is commonly used to compute individuals' probabilities of membership across race/ethnicity groups or to classify individuals into groups if probabilities exceed a certain threshold, with the probabilities or the single predicted group generally stored to facilitate further analysis. However, Meta, implemented a version of BISG where group estimations are made on the fly and immediately aggregated into summary statistics "[t]o avoid creating the conditions where individual-level race inferences would be generated."¹⁹⁴

Ephemeral methods may be particularly interesting to explore when it comes to observation and inference, where privacy risks and miscategorization/misrepresentation harms come into particular tension with fairness imperatives. However, data protection regimes that strictly limit the processing of sensitive characteristics do not necessarily differentiate between whether data is processed durably or ephemerally, leaving open questions about whether this sort of approach will comport with more stringent legal constraints in jurisdictions like California.

Privacy enhancing methods

Aggregation

Aggregation, or the combining and summarizing of data to reduce identifiability of individual data points, appears frequently in practitioner-described methods for fairness measurement. Aggregation can happen through *generalization*, or replacing data values with less precise values, as well as *suppression*, or removing outliers from datasets. While aggregation alone is a reasonably weak form of privacy protection, it can still help prevent inadvertent misuse and add friction to malicious reidentification efforts. K-anonymity, for instance, is an aggregation method to ensure that unique instances of data appear a minimum number of times in a dataset, making it more difficult to identify particular individuals based on uncommon constellations of data points.¹⁹⁵ The more records that are identical (or very similar) in a dataset, the less likely that people in the dataset who share those characteristics can be reidentified.

Airbnb, Meta, and Apple all describe incorporating aggregation into their demographic measurement methods. Airbnb describes using “ p -sensitive k -anonymity,” where datasets have a minimum number of distinct values within each subgroup (e.g. a dataset with a minimum of k identical values would have at least p distinct instances of a sensitive attribute within the set of identical values).¹⁹⁶ Similarly, Meta noted the use of k -anonymity in its BISG tool, designing the tool to only return aggregate metrics if a minimum number of people within the identified groups are represented. For example, if the tool were used to measure a population of 50 individuals, it would not return any results about a subgroup for which only two individuals were present in that population.¹⁹⁷ In a discussion of Apple’s fairness measurement work, the Partnership on AI described how a “secure aggregation protocol” might be employed, using cryptography to ensure that only aggregate demographic insights would be revealed to analysts.¹⁹⁸

Researchers have proposed a method called *quantification* to measure group characteristics without inferring individual data. This method involves using supervised learning to estimate the proportion of data points within a larger dataset that belongs to certain classes or groups, without assigning those characteristics to individual data points in the dataset.¹⁹⁹ In practice, this method resembles ephemerality since it involves calculating information about each data point and aggregating them on the fly to predict the demographic distribution of the measured population. This method does require some amount of labeled demographic data to train the estimation model, so would need to be coupled with other approaches like collection or auxiliary datasets.

Differential privacy

To address residual reidentification risk in aggregation and pseudonymization, researchers and practitioners have turned to statistical disclosure limitation techniques like suppressing uncommon values or swapping values with one another to reduce risk of disclosure for specific data points. Differential privacy offers more formal mathematical guarantees, involving the addition of a specific amount of random statistical noise to datasets to realize particular privacy constraints. Differential privacy introduces uncertainty to the fidelity of individual records in a dataset while preserving overall statistical properties of the dataset — in other words, the method “addresses the paradox of learning nothing about an individual while learning useful information about a population.”²⁰⁰ Differential privacy has been deployed in various contexts, from public health²⁰¹ to technology companies²⁰² to the US Census²⁰³ as a way to help protect against deidentification, and some practitioners have incorporated it into their fairness measurement work to help mitigate privacy risk. For instance, Apple incorporated differential privacy into its federated statistics efforts for post-deployment fairness measurement,²⁰⁴ and Meta included the method as part of its implementation of the BISG calculations that power its Variance Reduction System to address disparities in ad delivery.²⁰⁵

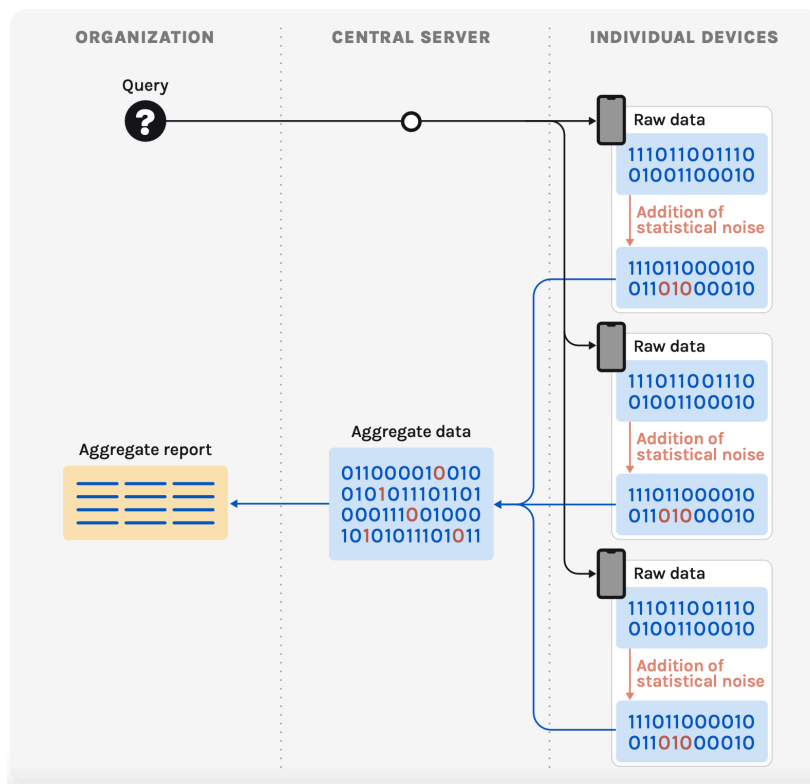


Figure 5. Local and Central Differential Privacy.

A graphic illustrating how statistical noise is added to data in different differential privacy paradigms.

Eyes Off My Data: Exploring Differentially Private Federated Statistics to Support Algorithmic Bias Assessments Across Demographic Groups, Partnership on AI, December 2023;

https://partnershiponai.org/wp-content/uploads/dlm_uploads/2023/12/PAI_whitepaper_eyes-off-my-data-1.pdf.

Differential privacy can be applied locally (where noise is added directly on a device or to an individual data point or batch) or globally (where noise is added to an overall dataset or once data reaches a central server). The choice between the two has implications for the amount of noise necessary to obscure individual data to the desired degree.²⁰⁶

It's important to note that since differential privacy involves adding statistical noise to a dataset, it does affect the fidelity of measurements, particularly for smaller populations and intersectional identities that may include even fewer people. The more noise (and resulting privacy protection) added, the less the measurements will reveal about true disparities. As such, the choice of the *epsilon* (ϵ) value, which represents the amount of statistical noise needed to uphold a certain privacy guarantee — sometimes called the privacy budget — has significant implications for the impact of differential privacy on measurement accuracy.²⁰⁷ (Across privacy enhancing approaches, this phenomenon is called the privacy-utility tradeoff, and remains an active area of research

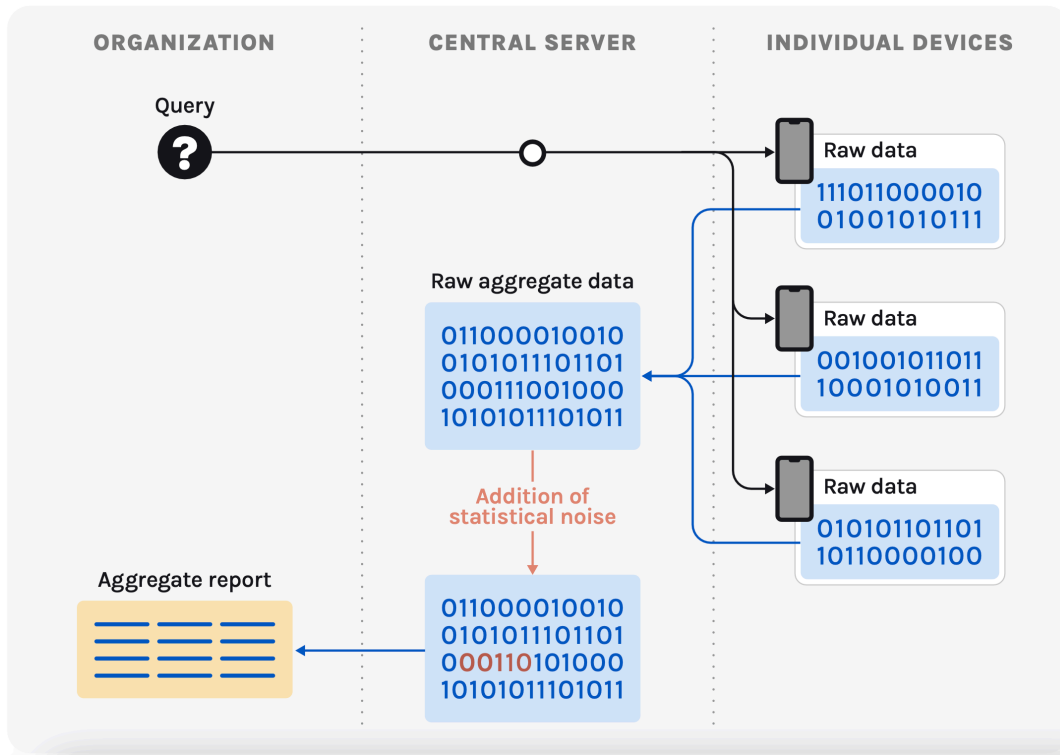


Figure 5 (continued). Local and Central Differential Privacy.

A graphic illustrating how statistical noise is added to data in different differential privacy paradigms.

Eyes Off My Data: Exploring Differentially Private Federated Statistics to Support Algorithmic Bias Assessments Across Demographic Groups, Partnership on AI, December 2023; https://partnershiponai.org/wp-content/uploads/dlm_uploads/2023/12/PAI_whitepaper_eyes-off-my-data-1.pdf.

and discussion.²⁰⁸) An ϵ value that provides an expected degree of privacy protection may lead to the exclusion of statistical minorities from analysis. For example, Meta explained that to enable their Variance Reduction System for mitigating racial disparities in ad delivery to function while still upholding differential privacy guarantees, the analysis required combining the census categories of Native American/Alaskan Native, Asian/Native Hawaiian/Pacific Islander, and Multiracial into a combined category.²⁰⁹

Keep in Mind

A fundamental challenge of differential privacy is the tradeoff between privacy protection and data utility: the stronger the mathematical privacy guarantees, the less informative data will be in reflecting real-world circumstances. This increased uncertainty can lead to misleading conclusions, including presumptions that no disparities exist when they actually do or vice versa.²¹⁰ Different communities also experience

disproportionate effects from the use of differential privacy. For example, smaller minoritized populations may require more statistical noise to be as protected from reidentification as the broader population. However, this additional noise can obscure disparities experienced by that community far more significantly than it does for larger populations.

While differential privacy does not by itself address demographic data-related harms like miscategorization or reinforcement of socially-defined categories, its contribution to protecting against data leakage can play an important role in a portfolio of efforts to strike a reasonable balance between privacy and equity needs. Practitioners should take care to select privacy parameters that provide sufficient protection to individuals represented in datasets while retaining the ability to detect with fidelity whether smaller communities may be facing disparities that warrant intervention. Direct engagement with both different communities and multiple representatives of each community is important in order to identify the right balance between privacy and data utility.²¹¹

Secure multi-party computation

Researchers have developed various cryptographic methods to enable institutions to conduct measurements over private datasets, in response to challenges and concerns around certain actors holding data about sensitive characteristics and constraints related to the sharing of such data between parties. One such method is secure multi-party computation (SMPC), which allows parties to conduct analyses across multiple datasets without sharing data with one another. SMPC has been both theorized and piloted as a method to enable demographic measurement while keeping demographic data contained to a trusted environment.

For example, some researchers have proposed a structure where a regulator would collect and hold encrypted versions of

users' sensitive data. Model developers could then use an SMPC protocol to conduct measurements using that data without directly accessing it, in order to certify whether a model conforms to certain fairness requirements.²¹² Meta piloted a version of this proposal, but instead of working with a regulator, they enrolled several third-party institutions to collect and cryptographically split demographic characteristics. As a result, no party held raw survey responses, but when securely combined through the SMPC protocol, the data could be used by the company for statistical fairness analysis.²¹³

Keep in Mind

While SMPC could be a promising technical solution for enabling measurement while avoiding the sharing of sensitive data about people across parties, it is technically complex and computationally costly. This means it may be challenging for smaller organizations or institutions without specialized expertise to take advantage of this method, tricky to explain to users in simple terms how data is safeguarded, and difficult for policymakers and regulators to govern. The complexity of privacy-enhancing technologies can also lead to implementation errors that, counterproductively, undermine privacy.²¹⁴ Moreover, this creative technical architecture also does not address key questions and limitations related to measuring the underlying data (discussed in Section 3), so must be considered as part of a portfolio of potential approaches to enable more responsible demographic measurement.

Procedural controls

User control

Discussions around sensitive data often focus on whether users have agency over the collection and use of details whose disclosure or misuse could be particularly harmful. Practitioners seeking to process demographic characteristics for fairness must navigate questions regarding how to inform people about how data will be used, obtain permission to process that data, and provide opportunity for data to be corrected or deleted. LinkedIn, for instance, clearly communicates to users that “you may update your information or remove it at any time by managing your settings. If you remove your data, we will delete it” (though this commitment appears to be specific to demographic data that users have directly provided, and not characteristics the company infers for aggregate insights or personalization like age or gender).²¹⁵ While providing clear **user control** appears to be less common in the context of observation and inference approaches, Airbnb gave users the opportunity to opt out of efforts to measure the effect of perceived race on host acceptance rates.²¹⁶ Meta offered users the choice of whether to participate in its race measurement survey, explaining to users why this data was being sought and how it would be used.

At the same time, like most other users of the BISG method, Meta did not describe offering users a choice in whether to be included in imputation-based analyses. Instead, Meta seems to have relied on several layers of privacy interventions to address potential concerns about the method and language in its privacy policy that it will process users’ data to “identify and combat disparities and racial bias against historically marginalized communities.”²¹⁷

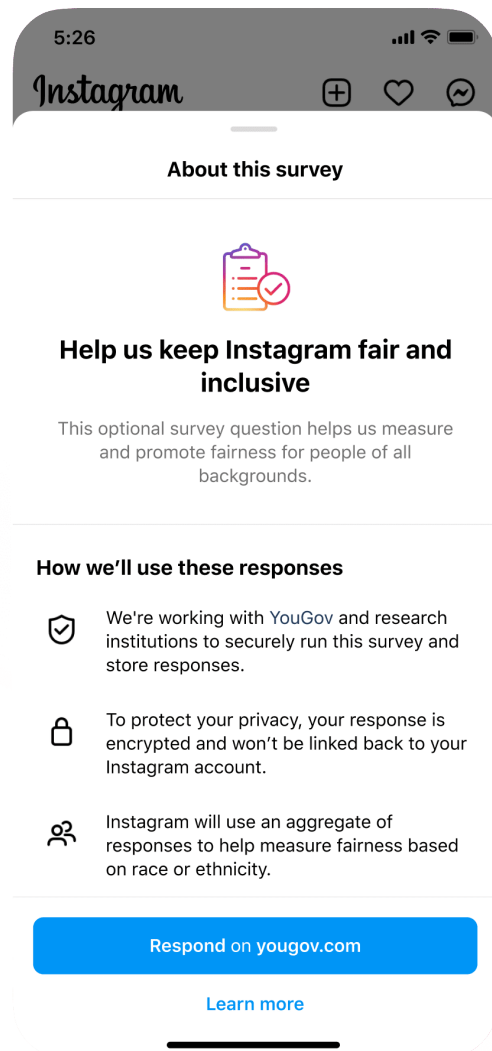
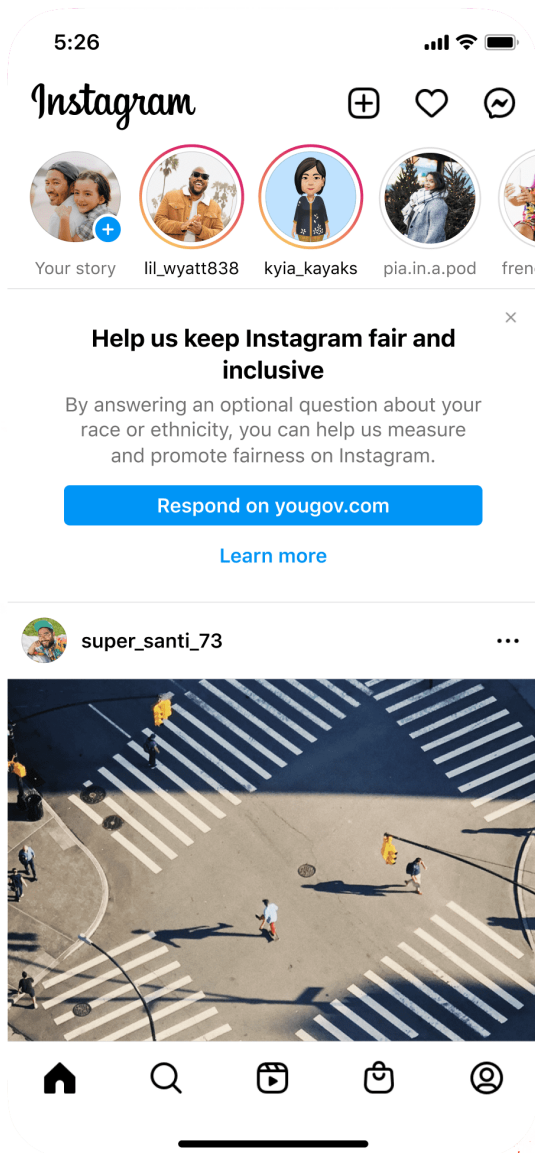


Figure 6. Race/ethnicity survey on Instagram.

Screenshots of prompts displayed to some Instagram users asking them for their race or ethnicity.

Instagram Blog, July 2022:

<https://about.instagram.com/blog/announcements/collecting-and-measuring-demographic-information>

Keep in Mind

Stakeholder discussions on demographic data hosted by the Partnership on AI reveal strong consensus that users should be invited to actively opt in to sharing data rather than provided an opportunity to opt-out.²¹⁸ However, the prevalence of observation and inference methods historically and currently used for civil rights enforcement raises questions about whether opt-in **consent** sufficiently enables systemic nondiscrimination efforts. The recently passed EU AI Act similarly acknowledged this limitation, clarifying that processing of special categories of personal data that is strictly necessary to ensure bias detection and correction in relation to high-risk AI systems is lawful under the legal basis of “substantial public interest,” not just consent.²¹⁹ While offering people agency over their identity data is critically important, there may be significant implications on the ability to detect disparities if participation rates vary significantly across communities. Advocates and regulators will need to grapple with this tension as requirements to proactively detect and mitigate biases solidify.

Organizational oversight

Once an organization has determined the method(s) it will use to seek out and handle demographic data, it needs to determine which individuals or teams may access and use that data and under what circumstances. Even with clear policies and hard-coded constraints, **organizational oversight** efforts may include cross-functional committees to review proposed uses of data or measurement methods to ensure they comply with those policies. These processes may also be used to make sure that relevant decision-makers, such as compliance or legal teams, have visibility into the measurements and their results to ensure that sufficient remediation actions are taken if needed. For example, Meta described that “analysts may only access BISG calculation tools

after receiving approval to do so through an internal, structured governance process; requests to access the relevant tooling will be reviewed by a committee including representatives from Meta's Civil Rights, Responsible AI, data science, policy and legal teams."²²⁰ The Partnership on AI's stakeholder conversations recommended that, in the context of differential privacy, institutions should also maintain organizational oversight of the process of querying data as well as the privacy parameters of those queries to ensure that privacy budgets are respected.²²¹ Board-level efforts to define organizational objectives, provide oversight related to operational risks, and authorize relevant initiatives may also play a role.

Separate teams

Regulated institutions commonly establish **separate teams** responsible for oversight and compliance; these teams typically also facilitate the process of restricting access to protected characteristic data to those whose job function requires it.²²² Such organization structures are particularly common in banking and financial institutions where responsibility for fair lending compliance is purposely housed in teams different from those building models and products. Airbnb has taken a similar approach, establishing an anti-discrimination product team that has exclusive access to perceived race data and a mandate to fight discrimination on the platform.²²³

Keep in Mind

While separate teams can be a useful approach to protect against data misuse, related decisions about organizational structure can have a marked impact on anti-discrimination efforts overall. First, specialized teams need to be sufficiently staffed both in terms of headcount and skill profiles to tackle known disparities and make proactive efforts to detect new ones.²²⁴ Second, smaller organizations may not have sufficient expertise across both modeling and compliance

teams to enable complete organizational separation; data scientists may support multiple teams and therefore permeate organizational firewalls.²²⁵ Finally, separation can have implications for how proactively and effectively teams developing products and models can search for and adopt less discriminatory alternatives. With separate teams, these searches tend to happen relatively late in the product or model development process when compliance teams are tasked with reviewing work against fairness and nondiscrimination requirements. This delay can have the perverse effect of measurement becoming a rote and narrow compliance exercise and constrain potential options to mitigate disparities that are detected. For example, if developer teams are unable to conduct measurements and compliance teams are unable to directly modify models, developer teams are unable to proactively train multiple model versions and easily compare them across protected characteristics or apply debiasing techniques in-line with model training. Enabling model development teams to conduct these searches directly could help systematize the identification of less discriminatory algorithms earlier, but would require a more substantial investment in oversight and technical controls to ensure data is not misused or inadvertently used in model training.²²⁶

Privacy impact assessments

Given the various considerations organizations face when working through options to measure and handle demographic data, thorough, cross-functional, and multistakeholder review of the possible and proposed approaches is essential to responsibly navigate the questions that inevitably arise related to demographic measurement. Structured data protection or **privacy impact assessments** can help organizations evaluate whether proposed methods for measuring and handling demographic data sufficiently mitigate against risks. These assessments are required for government agencies and industry actors in certain jurisdictions.²²⁷

Impact assessments should be more than perfunctory documentation exercises. They should involve representatives of communities and groups who may benefit from or be harmed by demographic data efforts to co-define methods of measurement, which social identity characteristics will be included, privacy/utility tradeoffs, definitions of fairness, and expectations around remediation.²²⁸ If communities have reason to believe that the definition of fairness an organization has selected will do more harm than good, for example, they may determine that the risks of sharing data in the first place outweighs the benefits.²²⁹ Airbnb navigated this by partnering with civil rights organization Color of Change and seeking input from a number of other advocacy groups to co-develop the company's approach.²³⁰ Apple worked with the Partnership on AI to convene two multidisciplinary expert workshops to reflect on the company's proposed measurement approach and surface recommendations. PAI documented that as part of that effort, "38 participant experts were drawn from a variety of backgrounds including industry, academic, and civil society experts specializing in racial, disability, and gender, and LGBTQIA+ equity, as well as data privacy and algorithmic fairness" to participate in three-hour workshops discussing the company's selected method and its limitations. Findings from the workshops were published in a lengthy public report.²³¹

Keep in Mind

While organizations may worry that disclosing early thinking about how to tackle bias measurement will open them to unwarranted critique, inviting external stakeholders into the process early, deeply, and consistently can help ensure the methods ultimately selected incorporate expected mitigation of privacy and other harms, and are recognized as legitimate by experts, regulators, and the public. Organizations should be mindful of who they involve in participatory exercises since decision outcomes will be shaped by whose voices are heard. Because communities are not homogeneous, including multiple representatives per group can help reveal divergent preferences.²³²

Transparency

Like all data processing and fairness efforts, **transparency** to both users and stakeholders plays a key role contributing to and demonstrating responsible data practices. If organizations are required or choose to disclose their data practices, they are less likely to engage in practices that would draw criticism or scrutiny. Conversely, if companies make false claims about their data practices, they may be liable for engaging in deception.²³³

Transparency can involve in-context disclosure to users about what data will be collected and how it will be used, public messaging like blog posts both about the organization's plans and the findings of its analyses, presentations at community meetings or town halls, and more thorough technical reports. (All examples in this report were able to be cited because information was shared publicly about them, though some cases were shared more prominently than others via official publication or information prominently displayed to users).

Keep in Mind

Experts have recommended that transparency cover not only what general data will be collected and analyzed but also how demographic groups are defined, the definition(s) of fairness that will be pursued and, when applicable, detailed privacy parameters that have been selected. Effective transparency also requires sound internal documentation. Practitioners should be as clear as possible about how they will and won't use people's data; to make sure these statements are accurate, strong internal controls like the ones described in the previous sections must be in place.

Thoughtful decisions around handling demographic data cannot address all harms involved in obtaining that data, particularly representational harms like miscategorization, misrepresentation, delegitimization, reinforcement of marginalization, or the calcification of administrative designations that have contributed to inequity in the first place. However, in cases where measurement is compelled or can meaningfully reduce blatant disparities, adoption of multiple handling safeguards is imperative to ensure these practices serve the goals of preventing discrimination and advancing equity rather than undermining them.



05

Discussion and Recommendations

As this report makes clear, the lack of immediately available raw demographic data should not stand in the way of efforts to measure discrimination or unfair disparities in AI systems.

Similarly, the lack of immediately available raw demographic data cannot and does not justify efforts to haphazardly collect and use such data. This report has detailed a number of creative approaches and modular safeguards that can be mixed and matched, and that can and have been used to gain insight to potential disparities, while also reducing risks.

As policymakers and practitioners build regulatory and technical infrastructure to make progress in this domain, we highlight several recommendations to ensure that the balance remains tipped toward beneficial measurement efforts:

Practitioners should:

- **Establish ongoing relationships with communities affected by measurement activities — particularly those who might experience disproportionate harm — to co-design data collection and handling strategies, discuss potential risks and benefits, and collaboratively define fairness goals.**

Meaningful engagement is core to ensuring measurement efforts avoid inadvertently harming already vulnerable communities, and should be sustained throughout the process of conducting measurement and mitigating identified gaps. Qualitative research approaches like ethnographic and participatory methods help add important nuance and incorporate lived experiences into research designs.

- **Where possible, consider methods that avoid collecting, generating, or storing sensitive demographic information in a way that can be easily connected to individuals.** Whether collected, inferred, or otherwise obtained, demographic data should be stored and accessed in as aggregate and deidentified manner as possible to avoid misuse or honeypots for malicious actors. Methods like surrogate characteristics, cohort discovery, and exploratory analysis can be useful to unlock actionable insights, while handling approaches like aggregation and ephemerality can help avoid reliance on individual data. If individual level data is necessary — for example, to provide remedy to those harmed or to comply with legal or regulatory requirements — technical and institutional friction should be added to ensure that data cannot be accessed or used for unauthorized purposes. Organizations should strive to strike a reasonable balance to enable the detection and mitigation of harms to people and communities while protecting them from other privacy, security, and representational harms, in consultation with affected stakeholders.
- **Take great care before using observation and inference methods to identify characteristics, especially those lacking precedent or that resist observation.** Practitioners should avoid pursuing new machine learning approaches to infer characteristics about individuals without clear, multi-stakeholder consensus, given the acute implications. Aggregate inference approaches that avoid generating individual prediction are a potentially promising path to explore, but deep, sustained engagement with communities must be integral to explorations in this domain.

- **Clearly differentiate between perceived or implied characteristics and actual ones**, and avoid reverting to shorthand that could undermine such distinctions when conducting observation and labeling of content. Care should also be taken to avoid using content-level or surrogate characteristics to assign characteristics to individuals.
- **Employ a robust combination of approaches to handling data and measurement methods to ensure appropriate use.** The choice of data and infrastructure controls, privacy enhancing methods, and procedural controls should be calibrated to the sensitivity of the data that is collected, generated, or stored, and take into account the risks and limitations presented by the measurement methods the organization has adopted.
- **Communicate openly about their demographic measurement efforts, as well as their handling of this data.** Users, communities, and stakeholders should be able to have a clear understanding of how a company approaches their demographic measurement efforts.

Government agencies and regulators should:

- **Recognize that a variety of approaches are available for companies to identify and measure disparities, even in the absence of comprehensive demographic data collection.** As such, agencies and regulators should expect organizations to make reasonable efforts to conduct algorithmic impact assessments and engage in non-discrimination efforts, particularly in consequential contexts.
- **Clarify criteria and expectations about acceptable measurement methods when it comes to civil rights compliance, and articulate minimum expectations for how data and methods should be handled.** The Consumer Financial Protection Bureau set an example by sharing details on its use of the BISG method to conduct fair lending analysis, establishing what has become a clear norm in the financial sector.²³⁴ Efforts toward this recommendation could include reconsidering existing prohibitions in civil rights law on the

collection of demographic data and enacting strong and nuanced privacy regulation to build trust that demographic measurement is possible without risk of data misuse.²³⁵ While some of the indirect measurement methods described in this report may be useful for detecting general patterns of disparity, their utility could be more complicated in the context of civil rights enforcement.

- **Explore and provide guidance on how measurement methods can be used to monitor compliance with Federal civil rights laws, including to conduct investigations and enforcement actions.** Direct measurement methods may not be available for all protected characteristics, but organizations should be expected to make reasonable efforts to understand potential biases within or driven by their relevant AI systems using one or more methods described in this report, with robust handling practices.
- **Facilitate collaboration between NGOs, research institutes, and government data agencies to explore creative ways that existing administrative data can be used to conduct measurements in a privacy-respecting manner.** Proposals such as the National Secure Data Service²³⁶ and emergent ideas like data trusts and data cooperatives may have some potential to unlock possibilities for auxiliary data without sharing people's data freely across institutions.²³⁷ However, setting up this sort of infrastructure will require in-depth consultations and sustained investment for its benefit to be realized.
- **Encourage and support continued research to explore how unsupervised, synthetic, privacy-enhancing, and content-related methods can be used to further the detection and remediation of bias and discrimination,** with a particular focus on how these methods can be used to advance nondiscrimination efforts in cases where labeled demographic data about individuals is unavailable, difficult to obtain, or presents too many risks to pursue safely.



06

Conclusion

“Respecting people’s self-determination and autonomy when it comes to sensitive data about who we are is complex and hard to do well. But ignoring that kind of data is also not an option.”²³⁸

~ Zara Rahman, Digital Rights Researcher

As governments and policymakers increasingly expect companies developing and using AI systems to proactively identify and mitigate bias and discrimination, navigating the foundational question of demographic measurement has become critically important. While there is no one-size-fits-all solution, this report makes clear that the lack of obvious access to raw demographic data should not be considered an insurmountable barrier to assessing AI systems for fairness, but neither should it provide a blanket justification for widespread or incautious data collection efforts. A variety of creative approaches and safeguards can be used to gain insight into disparate patterns while reducing the risks related to demographic measurement. From exploring privacy-preserving techniques to pursuing measurement of content-related bias when disparities affecting people are hard to measure directly, practitioners have a range of tools at their disposal.



While concrete measurement of disparities plays an important role in detecting, preventing and enforcing against discrimination, overly quantifying fairness work runs the risk of insufficiently engaging with the upstream sources of bias including structural and societal issues, and lead to interventions that only superficially address inequities while failing to engage with the root causes of those issues. In order for demographic measurement efforts to serve the goals of preventing discrimination and advancing equity rather than undermining them, practitioners must be intentional about the data and methods they adopt in order to prevent the exacerbation of vulnerabilities that marginalized communities already face when it comes to privacy, safety, and dignity. The methods described in this report can be useful to diagnose issues and inform efforts to address biases, but even systems that fare well in measurements could still lead to adverse effects, so mechanisms for accountability and redress will remain relevant and important.

We urge practitioners to move toward more proactive and systematic work to seek out and address fairness gaps in their products and services, but to also to integrate a broader lens that considers AI models, systems, and products in their broader societal context and consider creative ways to tackle the various ways in which bias manifests in this broader ecosystem.

As practitioners navigate this complex but important landscape, they should engage early and often with impacted communities, clearly document and communicate their practices, and embed strong technical and institutional safeguards to prevent misuse. Ultimately, responsible demographic measurement demands extraordinary care — for technical choices and their implications, but even more for the people and communities this work must ultimately serve.



07

Glossary

Aggregation

The combining and summarizing of data to reduce identifiability of individual data points. Aggregation can happen through generation (replacing data values with less precise values) or suppression (removing outliers).

Auxiliary datasets

External sources of demographic data that can be combined with or compared against existing data.

BI(F)SG

Bayesian Improved Surname Geocoding / Bayesian Improved First Name Surname Geocoding. A method to impute race and ethnicity of populations using a conditional combination of name and zip code.

Cohort discovery

Using manual or automatic pattern detection to look for clusters of people experiencing disproportionate errors or other negative outcomes.



Collection

Directly asking or giving individuals within a population the opportunity to share their identity characteristics

Comprehensive collection

The data collection practice where every individual within a specified population (e.g., applicants, users, students) is given the opportunity to share specific demographic details, usually via a structured form or survey.

Differential privacy

the addition of a specific amount of random statistical noise to datasets to realize particular privacy constraints. Differential privacy can be applied locally (directly on a device or to an individual data point or batch) or globally (at the level of an overall dataset or once data reaches a central server)

Encryption

Scrambling data so it can't be easily deciphered without a mathematical key

Ephemerality

Data matching or inferences computed on the fly without leading to the creation or storage of new data

Exploratory analysis

Reviewing a system's data, artifacts, and outcomes to reason about how the design, behavior, or other characteristics of the system might lead to negative impact for certain communities.

Handling

Methodological, technical, and organizational guardrails to help prevent data and methods from misuse.

Inference

Using signals correlated with demographic characteristics — via individual proxy characteristics, combinations of probabilities related to a handful of data points like surname and zip code, or more complex machine learning models — to predict the likelihood a person relates to a particular demographic characteristic. Also see observation and inference.

Infrastructural controls

Data and system architecture choices that limit how data and measurement methods can be access or used

Keywords and terms

Manually or automatically constructing lists of words and topics that relate to demographic characteristics and using them to probe systems.

Measuring

Approaches by which organizations may obtain, observe, access, impute, or otherwise understand demographic characteristics, approximations, or patterns

Observation and inference

Considering conspicuous traits to assign perceived demographic characteristics or otherwise predicting the likelihood of a person relating to a particular demographic group.

Observation and labeling

Automatically or manually assigning labels of phenotypical or readily apparent traits to unidentified people represented in audiovisual or text content.

Organizational oversight

Cross-functional committees or other processes to review proposed uses of data or measurement methods to ensure they comply with those policies, and to ensure relevant decision-makers, such as compliance or legal teams, have visibility into the measurements and their results to ensure that sufficient remediation actions are taken if needed

Partial collection

The data collection practice of asking a subset of the population of interest to self-provide demographic details in order to conduct statistical testing against the full population. This can include broad calls for response until a statistically significant and representative set of data has been collected, or constructing a specific sample group or panel.

People (lens of analysis)

Real, identifiable individuals with lived experiences, identities, and rights. (e.g. credit applicants who have been subject in whole or part to decisions informed by AI systems)

Privacy impact assessments

Structured impact assessments to evaluate whether proposed use of data, such as methods for measuring and handling demographic data, sufficiently mitigate against privacy risks

Privacy-enhanced observation or inference

Adding noise to or using aggregation within observation or inference procedures to address privacy concerns

Protected characteristics

Characteristics like age, gender, race, disability, religion, and others for which the law prohibits discrimination (sometimes called “protected classes”)

Proxies

See Proxies and Surrogate characteristics

Proxies and surrogate characteristics

Signals or features that correlate with certain demographic characteristics (such as geographic location, occupation, or primary language) that can be used to detect patterns or disparities without directly making assumptions about individuals’ demographics or making predictions based on these characteristics.

Pseudonymization

A de-identification technique that replaces personal identifiers with placeholder information or otherwise breaks the link between identifying data and other data about an individual

Qualitative research

Direct engagement with people using and affected by systems to capture more nuanced insights about people’s lived experience that quantitative methods may overlook (e.g. ethnography and participatory methods)

Representations (lens of analysis)

Depictions or references to people or identities that are not necessarily tied to specific individuals (e.g. unidentified faces present in photographs that have been cropped by an automated tool).

Retention

The period of time data is stored before it is deleted or destroyed, either through hard-coded rules or manual policies.

Secure multiparty computation

A cryptographic protocol that allows parties to conduct analyses across multiple datasets without sharing data with one another

Separate teams

Assigning a specific team to be responsible for oversight and compliance with laws that implicate demographic measurement, such as compliance teams

Surrogate characteristics

See Proxies and Surrogate characteristics

Synthetic data

Artificially generated data that simulates the structure and distribution of real-world examples or populations.

Transparency

Disclosure about data measurement and handling practices and other relevant information (e.g. definitions of fairness) relevant to bias and equity measurement efforts, including in-context disclosure, public communications, and in-depth public documentation

User control

Providing people with the opportunity to decide whether to share their data and to request data be corrected or deleted



08

Endnotes

- 1 Zara Rahman, *Who Do They Think You Are? Categories, Classification, and Profiling*, SSRN, <https://just-tech.ssrc.org/articles/who-do-they-think-you-are-categories-classification-and-profiling> [perma.cc/TN82-TB27]
- 2 Blueprint for an AI Bill of Rights: Making Automated Systems Work for the American People, White House (2022), <https://www.whitehouse.gov/ostp/ai-bill-of-rights> [<https://perma.cc/74NG-TVAK>]; Executive Order on Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence, Executive Office of the President, (2023), <https://www.whitehouse.gov/briefing-room/presidential-actions/2023/10/30/executive-order-on-the-safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence> [<https://perma.cc/SB7L-P5W9>]; Statement of Commissioner Alvaro M. Bedoya On *FTC v. Rite Aid Corporation & Rite Aid Headquarters Corporation* (2023), https://www.ftc.gov/system/files/ftc_gov/pdf/2023190_commissioner_bedoya_riteaid_statement.pdf. [<https://perma.cc/74QA-9SJ2>]
- 3 Miranda Bogen, Aaron Rieke, and Shazeda Ahmed, *Awareness in Practice: Tensions in Access to Sensitive Attribute Data for Antidiscrimination*, Conference on Fairness, Accountability, and Transparency (FAT* '20), <https://doi.org/10.1145/3351095.3372877> [<https://perma.cc/PQW3-7E8V>]; McKane Andrus, Elena Spitzer, Jeffrey Brown, and Alice Xiang, *What we can't measure, we can't understand: Challenges to demographic data procurement in the pursuit of fairness*, Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT '21), <https://doi.org/10.1145/3442188.3445888> [<http://perma.cc/WK5V-C87T>].
- 4 McKane Andrus and Sarah Villeneuve, *Demographic-Reliant Algorithmic Fairness: Characterizing the Risks of Demographic Data Collection in the Pursuit of Fairness*, 2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT '22), <https://doi.org/10.1145/3531146.3533226> [<https://perma.cc/AR5M-PKVX>]; William Seltzer and Margo Anderson, *The Dark Side of Numbers: The Role of Population Data Systems in Human Rights Abuses*, Social Research (2001), <https://www.proquest.com/docview/209667997>. [<https://perma.cc/A7A8-HERH>]
- 5 Jennifer King, Daniel Ho, Arushi Gupta, Victor Wu, Helen Webley-Brown, *The Privacy-Bias Tradeoff: Data Minimization and Racial Disparity Assessments in U.S. Government*, Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency (FAccT '23), <https://dl.acm.org/doi/10.1145/3593013.3594015>. [perma.cc/KRT3-YQQC].

- 6 Carolyn Ashurst and Adrian Weller, *Fairness Without Demographic Data: A Survey of Approaches, Equity and Access in Algorithms, Mechanisms, and Optimization* EAAMO '23, <https://doi.org/10.1145/3617694.3623234>. [perma.cc/RL87-KMNJ]
- 7 *How the student debt crisis affects African Americans and Latinos*, Washington Center for Equitable Growth (2016), <https://equitablegrowth.org/how-the-student-debt-crisis-affects-african-americans-and-latinos>. [perma.cc/68ZA-7TX9]
- 8 "OpenAI's GPT Is a Recruiter's Dream Tool. Tests show there's racial Bias," Bloomberg (March 2024), <https://www.bloomberg.com/graphics/2024-openai-gpt-hiring-racial-discrimination>.
- 9 Lauren Davenport, *The Fluidity of Racial Classifications*, Annual Review of Political Science (2019), <https://doi.org/10.1146/annurev-polisci-060418-042801>. [perma.cc/3SQ5-ZHSU]
- 10 Rachel Marks and Nicholas Jones, "Collecting and Tabulating Ethnicity and Race Responses in the 2020 Census," U.S. Census Bureau (February 2020), <https://www2.census.gov/about/training-workshops/2020/2020-02-19-pop-presentation.pdf>. [perma.cc/4AQR-P4KN]
- 11 Civil rights groups have explained that many Latine people do self-identify with race categories when race and ethnicity are separated, which can hinder the detection of disparities. Fact sheet: *Why do we need a combined race and ethnicity question?*, Leadership Conference on Civil and Human Rights (April 2023), <https://civilrights.org/wp-content/uploads/2023/04/FAQCombinedQuestion-1.pdf> [<https://perma.cc/XJC9-T2BP>]. The US recently announced that it would combine the race and ethnicity categories and add a Middle Eastern and North African category. Mike Schneider "US changes how it categorizes people by race and ethnicity. It's the first revision in 27 years," *Associated Press* (March 2024), <https://apnews.com/article/race-ethnicity-census-bureau-hispanics-0b2c325b683efd95e8e8e24235654abd>. [<https://perma.cc/25HB-FRTX>]
- 12 Hansi Lo Wang, "The U.S. census sees Middle Eastern and North African people as white. Many don't," NPR (February 17, 2022), <https://www.npr.org/2022/02/17/1079181478/us-census-middle-eastern-white-north-african-mena> [<https://perma.cc/8LPC-ZH44>]. The Biden administration has proposed adding this category (and combining the race and ethnicity categories) for the 2030 census. Hansi Lo Wang, "New 'Latino' and 'Middle Eastern or North African' checkboxes proposed for U.S. forms," NPR (April 2023), <https://www.npr.org/2023/01/26/1151608403/mena-race-categories-us-census-middle-eastern-latino-hispanic>. [<https://perma.cc/QEP7-GHUT>]
- 13 *Draft Proposed New Regulation Concerning Quantitative Testing of External Consumer Data and Information Sources, Algorithms, and Predictive Models Used for Life Insurance Underwriting for Unfairly Discriminatory Outcomes*, Colorado Division of Insurance (2023), <https://drive.google.com/file/d/1BMFuRKbh39Q7YckPqrhrCRuWp29vJ44O/view> [perma.cc/V55B-DS55]

- 14 "Census Bureau Releases 2020 Census Population for More Than 200 New Detailed Race and Ethnicity Groups," US Census Bureau (September 2023), <https://www.census.gov/library/stories/2023/09/2020-census-dhc-a-race-overview.html> [perma.cc/ZN42-NM2G]
- 15 Year 1 Report, National Artificial Intelligence Advisory Committee (NAIAC) (2023), <https://www.ai.gov/wp-content/uploads/2023/05/NAIAC-Report-Year1.pdf>. [perma.cc/5GSD-9V3W]
- 16 Department of Education, *Final Guidance on Maintaining, Collecting, and Reporting Racial and Ethnic Data to the U.S. Department of Education* (2007), <https://www.govinfo.gov/content/pkg/FR-2007-10-19/pdf/E7-20613.pdf> [perma.cc/E3UF-JTTC]
- 17 Healthcare entities seeking reimbursement from Medicare and Medicaid programs do have compliance obligations under Title VI of the Civil Rights Act. but data collection is not explicitly required under the act. *National Research Council (US) Panel on DHHS Collection of Race and Ethnic Data*, National Academies Press (2004), <https://www.ncbi.nlm.nih.gov/books/NBK215758>. [perma.cc/Q2JF-V87N]
- 18 LinkedIn Settings, <https://www.linkedin.com/mypreferences/d/demographic-info> (accessed February 20, 2024) [perma.cc/WVC8-9LGZ]
- 19 Monica Lewis, "How We Can All Help Create Equal Access to Opportunity," LinkedIn Official Blog (February 2021), <https://www.linkedin.com/blog/member/equity/how-we-can-all-help-create-equal-access-to-opportunity>. [perma.cc/5C78-CG3L]
- 20 Johanna Wright, "Updates on our efforts to make YouTube a more inclusive platform," YouTube Official Blog (December 2020), <https://blog.youtube/news-and-events/make-youtube-more-inclusive-platform>. [perma.cc/6K28-A46Y]
- 21 "Understanding our neighbors to address bias and provide a safe and welcoming platform for all," Nextdoor Blog (April 2023), <https://blog.nextdoor.com/2020/06/11/a-nextdoor-where-everyone-belongs/#bias>.
- 22 Meta, constructed a purpose-built dataset to measure performance disparities in computer vision and audio AI models by collecting scripted video samples from several thousand research participants, alongside self-provided demographic characteristics including age, gender, language, disability status, and physical attributes. This dataset could then be used to test AI models even if the research participants who provided their demographic data were not present in models' initial training data. Bilal Porgali, Vítor Albiero, Jordan Ryda, Cristian Canton Ferrer, and Caner Hazirbas, *The Casual Conversations v2 Dataset*, arXiv (2023), <https://ai.meta.com/research/publications/the-casual-conversations-v2-dataset>. [https://perma.cc/YY5Y-6JNY]
- 23 Instagram launched a voluntary, off-platform survey inviting users to share their race/ethnicity. "Understanding How Different Communities Experience Instagram," Instagram Blog (July 2022), <https://about.instagram.com/blog/announcements/collecting-and-measuring-demographic-information>.

- 24 Pymetrics asked players to take an optional survey to self-identify and used responses to construct hold-out test sets (reported 75% response rates which is abnormally high). Christo Wilson, Avijit Ghosh, Shan Jiang, Alan Mislove, Lewis Baker, Janelle Szary, Kelly Trindel, and Frida Polli, *Building and Auditing Fair Algorithms: A Case Study in Candidate Screening*, Conference on Fairness, Accountability, and Transparency (FAccT '21), <https://doi.org/10.1145/3442188.3445928>. [perma.cc/A429-FW8P]
- 25 Zara Rahman, *Who Do They Think You Are? Categories, Classification, and Profiling*, SSRIC, <https://just-tech.ssrc.org/articles/who-do-they-think-you-are-categories-classification-and-profiling>. [perma.cc/2M4K-QU53]
- 26 Eburn Joseph, *Race: the unmarked marker in racialised hierarchical social systems*, Critical Race Theory and Inequality in the Labour Market: Racial Stratification in Ireland (Manchester 2020), <https://doi.org/10.7765/9781526134400.00007>. [perma.cc/Z7NF-YKTW]
- 27 For example, according to the United Nations Convention on the Rights of People with Disabilities, disability is not a static characteristic but an evolving state defined by the interaction between people with impairments, people with whom they interact, and environmental barriers. Shari Trewin, *AI Fairness for People with Disabilities: Point of View*, IBM Accessibility Research (2018), <https://arxiv.org/pdf/1811.10670.pdf>. [perma.cc/KQZ2-YNCC]
- 28 *Demographic-Reliant Algorithmic Fairness*
- 29 Charlotte Vuyiswa McClain-Nhlapo and Jenny Lay-Flurrie, "Narrowing the data gap: World Bank and Microsoft commit to unlocking better development outcomes for persons with disabilities," Official Microsoft Blog (June 2022), <https://blogs.microsoft.com/blog/2022/06/15/narrowing-the-data-gap-world-bank-and-microsoft-commit-to-unlocking-better-development-outcomes-for-persons-with-disabilities>. [perma.cc/F66A-8ZLG]
- 30 Joon Sung Park, Ece Kamar, Danielle Bragg, and Meredith Ringel Morris, *Designing an Online Infrastructure for Collecting AI Data From People With Disabilities*, Conference on Fairness, Accountability, and Transparency (FAccT '21), <https://dl.acm.org/doi/10.1145/3442188.3445870>. [perma.cc/757U-APVM]
- 31 White House, *Recommendations On The Best Practices For The Collection Of Sexual Orientation And Gender Identity Data On Federal Statistical Surveys* (2023), <https://www.whitehouse.gov/wp-content/uploads/2023/01/SOGI-Best-Practices.pdf>. [perma.cc/7BBE-29N7]
- 32 For an in-depth discussion of this challenge in the context of disability, see Nenad Tomasev, Kevin R. McKee, Jackie Kay, and Shakir Mohamed, *Fairness for Unobserved Characteristics: Insights from Technological Impacts on Queer Communities*, Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society (AIES '21), <https://doi.org/10.1145/3461702.3462540>. [https://perma.cc/NG74-PFZH]

- 33 *Automated Employment Decision Tools: Frequently Asked Questions*, NYC Department of Consumer and Worker Protection, <https://www.nyc.gov/assets/dca/downloads/pdf/about/DCWP-AEDT-FAQ.pdf> [perma.cc/C9CN-QVFJ]
- 34 E.g. “New Coalition Addresses Health Equity Challenges through Inclusive Data Collection,” Blue Cross Blue Shield Association (September 2023), <https://www.bcbs.com/press-releases/new-coalition-addresses-health-equity-challenges-through-inclusive-data-collection> [https://perma.cc/S5F9-KTYB]; *Improving Data on Race and Ethnicity: A Roadmap to Measure and Advance Health Equity*, Grantmakers In Health (GIH) and the National Committee for Quality Assurance (NCQA) (2021), <https://www.ncqa.org/wp-content/uploads/2022/01/GIH-Commonwealth-Fund-federal-data-report-part-2-1.pdf>. [https://perma.cc/TBY4-7WJ7]
- 35 Imani Dunbar, “It’s OK Not To Self-ID on LinkedIn, but Here Is Why You Should,” LinkedIn (September 2021), <https://www.linkedin.com/pulse/its-ok-self-id-linkedin-here-why-you-should-imani-dunbar> [https://perma.cc/DSZ7-XX2Z]; Ryan Roslansky, “Driving Equitable Outcomes: A Journey We’re Taking Together,” LinkedIn Official Blog (February 2021), <https://www.linkedin.com/blog/member/equity/driving-equitable-outcomes-a-journey-we-re-taking-together>. [https://perma.cc/WN3C-RBDV]
- 36 @mosseri, Instagram, July 8, 2022, <https://www.instagram.com/reel/CgjqD8bgMnC/>.
- 37 The UK’s Center for Data Ethics and Innovation has noted that “[d]ata collected directly from service users is likely to contain at least a degree of inaccuracy due to some users accidentally or intentionally misreporting their demographic traits.” *Report: Enabling responsible access to demographic data to make AI systems fairer*, Centre for Data Ethics and Innovation, Department for Science, Innovation & Technology (June 2023), <https://www.gov.uk/government/publications/enabling-responsible-access-to-demographic-data-to-make-ai-systems-fairer/report-enabling-responsible-access-to-demographic-data-to-make-ai-systems-fairer>. [https://perma.cc/H78U-ATDE]. This may be particularly prevalent for populations whose characteristics are not physically apparent (e.g. trans identities) and for which there is some stigma attached (e.g. mental health conditions). See e.g. *Measuring Health Inequities when Administrative or Self-Reported Race/Ethnicity Data is Incomplete*, Marc N. Elliott, National Network Workshop, UCLA (October 2021), <https://healthpolicy.ucla.edu/our-work/training/imputation-methods-increasing-raciaethnic-data-disaggregation> [https://perma.cc/Q85F-QVY7] (explaining that self-reported data can have high, non-random missingness).

- 38 Sarah Villeneuve and McKane Andrus, "Knowing the Risks: A Necessary Step to Using Demographic Data for Algorithmic Fairness," Partnership on AI Blog (September 2021), <https://partnershiponai.org/demo-data-knowing-the-risk> [perma.cc/F52L-Q46W]. While there are protections in some jurisdictions against discrimination or retaliation for disclosing protected characteristics, those protections can be under-enforced; even where they exist, people may not know about them and the protections they should confer. Moreover, there are myriad examples where disclosure has led to unlawful termination or other detrimental outcomes. See e.g. Grace Dean, "Papa John's hired a blind worker, then fired him before he'd worked a shift when he asked to bring his service dog to work, a federal agency says," Business Insider (December 2023), <https://www.businessinsider.com/papa-johns-fired-blind-worker-bring-service-dog-work-eeoc-2023-12> [perma.cc/PXD9-QPUF].
- 39 E.g. Brian W. Roberts, Jady Yao, Christian J. Trzeciak et al, *Income Disparities and Nonresponse Bias in Surveys of Patient Experience*, Journal of General Internal Medicine (2020), <https://doi.org/10.1007/s11606-020-05677-6> [perma.cc/UZ9F-JZKJ]; *Report: Enabling responsible access to demographic data to make AI systems fairer*.
- 40 E.g. Mike Schneider, "The Census Bureau wants to change how it asks about disabilities. Some advocates don't like it," Associated Press (December 2023), <https://apnews.com/article/disability-census-covid19-survey-count-backlash-18678c34ca19e66876faf7dcbdb86f5>. [perma.cc/4RHL-EDMJ]
- 41 Jessica Lynne Hicksted, *Stigma Associated with Invisible Disabilities and Its Effect on Intended Disclosure in the Workplace*, Walden Dissertations and Doctoral Studies (2023), <https://scholarworks.waldenu.edu/dissertations/12126> [https://perma.cc/54MZ-V9QL]; Pooja Jain-Link and Julia Taylor Kennedy, "Why People Hide Their Disabilities at Work," Harvard Business Review (June 2019), <https://hbr.org/2019/06/why-people-hide-their-disabilities-at-work>. [https://perma.cc/34R7-ECHG]
- 42 *Awareness in Practice*
- 43 *The Casual Conversations v2 Dataset*
- 44 Joon Sung Park, Danielle Bragg, Ece Kamar, and Meredith Ringel Morris, *Designing an Online Infrastructure for Collecting AI Data From People With Disabilities*, Conference on Fairness, Accountability, and Transparency (FAccT '21), <https://doi.org/10.1145/3442188.3445870> [perma.cc/CE5W-F8VA]
- 45 Julia Carrie Wong, "Google reportedly targeted people with 'dark skin' to improve facial recognition," The Guardian (October 2019), <https://www.theguardian.com/technology/2019/oct/03/google-data-harvesting-facial-recognition-people-of-color>. [perma.cc/X4HR-R7EE]
- 46 Donavyn Coffey, "Māori are trying to save their language from Big Tech," Wired (April 2021), <https://www.wired.co.uk/article/maori-language-tech>.

- 47 *E.g. Marc N Elliott, Allen Fremont, and Peter A Morrison et al, A new method for estimating race/ethnicity and associated disparities where administrative records lack self-reported race/ethnicity*, Health Services Research (2008), <https://pubmed.ncbi.nlm.nih.gov/18479410> [<https://perma.cc/5YW8-BZEG>] and Melony E. Sorbero, Roald Euller, Aaron Kofner, and Marc N. Elliott, *Imputation of Race and Ethnicity in Health Insurance Marketplace Enrollment Data, 2015–2022 Open Enrollment Periods*, RAND (2022), <https://aspe.hhs.gov/sites/default/files/documents/931e1e3edec745ea7fdd364f9e28d6c6/aspe-rand-imputation-race-ethnicity-mktplc-rpt.pdf>. [<https://perma.cc/33R8-4ZZZ>]
- 48 *Draft Proposed New Regulation Concerning Quantitative Testing of External Consumer Data and Information Sources, Algorithms, and Predictive Models Used for Life Insurance Underwriting for Unfairly Discriminatory Outcomes*, Colorado Division of Insurance (2023), <https://drive.google.com/file/d/1BMFuRKbh39Q7YckPqrhrCRuWp29vJ44O/view>. [perma.cc/7Q69-P22J]
- 49 See e.g. *Fair Lending Monitorship of Upstart Network’s Lending Model: Initial Report of the Independent Monitor*, Relman Colfax PLLC (2021), https://www.relmanlaw.com/media/cases/1088_Upstart%20Initial%20Report%20-%20Final.pdf [perma.cc/AX9H-VZD5]
- 50 “Inferred Age or Gender on LinkedIn,” <https://www.linkedin.com/help/linkedin/answer/a517610> (accessed March 2024). [perma.cc/FH54-XX3D]
- 51 While perceived race was generated through observation by the research partner, users were given the opportunity to opt-out. “A new way we’re fighting discrimination on Airbnb,” Airbnb Resource Center, (June 2020), <https://www.airbnb.com/resources/hosting-homes/a/a-new-way-were-fighting-discrimination-on-airbnb-201>. [perma.cc/N4WG-ARRS]
- 52 *A Report to Uber Technologies, Inc. On Its Efforts To Promote Civil Rights, Diversity, Equity, and Inclusion*, Covington (2023), https://s23.q4cdn.com/407969754/files/doc_governance/2023/Uber-CRA-Report-August-2023.pdf. [perma.cc/YCC8-GN52]
- 53 Specific interventions include obfuscating race/ethnicity categories using encryption at run-time, aggregating encrypted labels prior to so that decrypted results only display aggregate statistics, and incorporating differential privacy into measurements to minimize reidentification risk. “How Meta is working to assess fairness in relation to race in the U.S. across its products and systems”; Roy L. Austin Jr, “An Update on Our Ads Fairness Efforts,” Meta Newsroom (January 2023), <https://about.fb.com/news/2023/01/an-update-on-our-ads-fairness-efforts/> [perma.cc/3S5H-4A56]
- 54 *Report: Enabling responsible access to demographic data to make AI systems fairer*
- 55 Joseph P. Near and David Darais, *Guidelines for Evaluating Differential Privacy Guarantees*, NIST (2023), <https://doi.org/10.6028/NIST.SP.800-226.ipd>.

- 56 Brian Asquith, Brad Hershbein, and Tracy Kugler et al, *Assessing the Impact of Differential Privacy on Measures of Population and Racial Residential Segregation*, Harvard Data Science Review, <https://doi.org/10.1162/99608f92.5cd8024e>. [perma.cc/2ACT-K78A]
- 57 David R. Nerenz, Rita Carreon, and German Veselovskiy, *Race, ethnicity, and language data collection by health plans: findings from 2010 AHIPF-RWJF survey*, Journal of Health Care for the Poor and Underserved, <https://muse.jhu.edu/article/524354/pdf>. [perma.cc/TGJ7-C6WB]
- 58 *Defining Categorization Needs for Race and Ethnicity Data*, Race, Ethnicity, and Language Data: Standardization for Health Care Quality Improvement, Agency for Healthcare Quality and Research (2019), <https://www.ahrq.gov/research/findings/final-reports/iomracereport/reldata3.html>. [https://perma.cc/Q4UN-5HBK]
- 59 Hadi Elzayn, and Emily Black et al, *Estimating and Implementing Conventional Fairness Metrics With Probabilistic Protected Features*, arXiv (2023), <https://doi.org/10.48550/arXiv.2310.01679>. [perma.cc/ZL9K-VGYJ]
- 60 As the company explained “Unlike more common uses of BISG where full legal names may be consistently available to analysts, users of social media platforms may choose to register using the name they go by in real life. Meanwhile, imprecise location data that companies may opt to use to predict home (or primary) location is often derived, and people’s choices around whether to grant platforms permission to use precise location information can affect the accuracy of those predictions” whereas traditional uses of BISG use self-provided home address data. [*How Meta is working to assess fairness in relation to race in the U.S. across its products and systems.*](#)
- 61 The US Department of Agriculture adjusted its policy in the context of food assistance services in 2021, prohibiting the use of observed data in favor of self-identified data and auxiliary data such as census statistics and school enrollment data, and codified the new policy as a formal rule in 2023. *Questions and Answers Related to Collection of Race and Ethnicity Data by Visual Observation and Identification in the Child and Adult Care Food Program and Summer Food Service Program - Policy Rescission*, United States Department of Agriculture (2022), <https://www.cacfp.org/assets/pdf/CACFP+Collection+of+Race+and+Ethnicity+Data+by+Visual+Observation+and+ID+Policy+Rescission> [https://perma.cc/K8VL-ZDGE]; Food and Nutrition Service, *Supplemental Nutrition Assistance Program: Revision of Civil Rights Data Collection Methods* (2023), <https://www.federalregister.gov/documents/2023/12/14/2023-27351/supplemental-nutrition-assistance-program-revision-of-civil-rights-data-collection-methods>. [https://perma.cc/535Y-DZNA]
- 62 Fernanda CG Polubriaginof, Patrick Ryan and Hojjat Salmasian et al, *Challenges with quality of race and ethnicity data in observational databases*, Journal of the American Medical Informatics Association: JAMIA, <https://doi.org/10.1093/jamia/ocz113>. [perma.cc/C5VJ-PESX]

- 63 *Standards for the Classification of Federal Data on Race and Ethnicity*, Office of Management and Budget (1995), https://obamawhitehouse.archives.gov/omb/fedreg_race-ethnicity [<https://perma.cc/6MZA-V4XG>]; Dulce Gonzalez, Nancy López, and Michael Karpman et al, *Observing Race and Ethnicity through a New Lens An Exploratory Analysis of Different Approaches to Measuring 'Street Race'*, Urban Institute (2022), <https://www.urban.org/sites/default/files/2022-12/Observing%20Race%20and%20Ethnicity%20through%20a%20New%20Lens.pdf>. [<https://perma.cc/EM66-BTR4>]
- 64 Nathan Kallus, Xiaojie Mao, and Angela Zhou, *Assessing Algorithmic Fairness with Unobserved Protected Class Using Data Combination*, Management Science, <https://pubsonline.informs.org/doi/10.1287/mnsc.2020.3850>. [<https://perma.cc/FCA6-BK5X>]
- 65 Jiahao Chen, Nathan Kallus, and Xiaojie Mao et al, *Fairness Under Unawareness: Assessing Disparity When Protected Class Is Unobserved*, Conference on Fairness, Accountability, and Transparency (FAT* '19), <https://doi.org/10.1145/3287560.3287594>. [perma.cc/4UYW-JLPS]
- 66 *Demographic-Reliant Algorithmic Fairness; Fairness Under Unawareness: Assessing Disparity When Protected Class Is Unobserved*
- 67 Aaron Rieke, Vincent Southerland, Dan Svirsky, and Mingwei Hsu, *Imperfect Inferences: A Practical Assessment*, Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT '22), <https://doi.org/10.1145/3531146.3533140>. [perma.cc/5WBW-TGT7]
- 68 Solvejg Wastvedt, Joshua Snoke, and Denis Agniel et al, *De-Biasing the Bias: Methods for Improving Disparity Assessments with Noisy Group Measurements*, arXiv, <https://doi.org/10.48550/arXiv.2402.13391> [<https://perma.cc/C6EU-7AV2>]; Benjamin Lu, Jia Wan, Derek Ouyang, Jacob Goldin, and Daniel E. Ho, *Quantifying the Uncertainty of Imputed Demographic Disparity Estimates: The Dual-Bootstrap*, NBER (2024), https://conference.nber.org/conf_papers/f182298.pdf. [<https://perma.cc/V82U-LJYW>]
- 69 Ioan Voicu, *Using First Name Information to Improve Race and Ethnicity Classification* (2016), <http://dx.doi.org/10.2139/ssrn.2763826>. [perma.cc/44CW-W2YY]
- 70 Kosuke Imai et al, *Addressing census data problems in race imputation via fully Bayesian Improved Surname Geocoding and name supplements*, Science Advances (2022), <https://www.science.org/doi/10.1126/sciadv.adc9824>. [<https://perma.cc/7R28-DB73>]
- 71 Cory McCartan, Jacob Goldin, Daniel E. Ho, and Kosuke Imai, *Estimating Racial Disparities When Race is Not Observed*, arXiv (2023), <https://arxiv.org/pdf/2303.02580.pdf> [perma.cc/WML9-2EQV]
- 72 Kasey Matthews, "Improving this algorithm can make lending a lot less racist," Zest.ai (August 2020), <https://www.zest.ai/insights/improving-this-algorithm-can-make-lending-a-lot-less-racist> [<https://perma.cc/D2HQ-6JWY>]; "Zest Race Predictor," <https://github.com/zestai/zrp> [<https://perma.cc/Y7F9-7U4L>] (access March 1, 2024).

- 73 Lingwei Cheng and Isabel O. Gallegos et al, *How Redundant are Redundant Encodings? Blindness in the Wild and Racial Disparity when Race is Unobserved*, 2023 ACM Conference on Fairness, Accountability, and Transparency (FAcCT '23), <https://doi.org/10.1145/3593013.3594034>. [perma.cc/Z7RV-529P]
- 74 *Demographic-Reliant Algorithmic Fairness; Os Keyes, The Misgendering Machines: Trans/HCI Implications of Automatic Gender Recognition*, Proceedings of the ACM on Human-Computer Interaction (CSCW), <https://doi.org/10.1145/3274357>. [perma.cc/R2UP-QSXM]
- 75 *Report: Enabling responsible access to demographic data to make AI systems fairer*
- 76 *Fairness for Unobserved Characteristics: Insights from Technological Impacts on Queer Communities*; Trenton W. Ford, "Is your face gay? Conservative? Criminal? AI researchers are asking the wrong questions," Bulletin of the Atomic Scientists (May 2022), <https://thebulletin.org/2022/05/is-your-face-gay-conservative-criminal-ai-researchers-are-asking-the-wrong-questions/>. [perma.cc/F4HS-PV3D]
- 77 European Union Artificial Intelligence Act ("EU AI Act"), https://www.europarl.europa.eu/doceo/document/TA-9-2024-0138-FNL-COR01_EN.pdf. [https://perma.cc/527H-9MMC]
- 78 Vinicia Perkins, *The Illusion of French Inclusion: The Constitutional Stratification of French Ethnic Minorities*, Georgetown Journal of Law and Modern Critical Race Perspectives (February 2019), <https://www.law.georgetown.edu/mcrp-journal/wp-content/uploads/sites/22/2019/12/GT-GCRP190031.pdf>. [perma.cc/746S-CP55]
- 79 *Using publicly available information to proxy for unidentified race and ethnicity: A methodology and assessment*, Consumer Financial Protection Bureau (2014), https://files.consumerfinance.gov/f/201409_cfpb_report_proxy-methodology.pdf [perma.cc/H4XC-QTWB]; *Fairness Under Unawareness: Assessing Disparity When Protected Class Is Unobserved*.
- 80 In a compliance metrics report per DOJ's settlement with Meta, the independent reviewer found the company's ads fairness metrics to be sensitive to the selection of this threshold, but noted it was consistent with academic, industry, and regulatory literature. *VRS Compliance Metrics Verification*, Guidehouse (June 2023), <https://www.justice.gov/usao-sdny/file/1306631/dl?inline>. [perma.cc/SMV6-W5A6]
- 81 *Climate and Economic Justice Screening Tool: Frequently Asked Questions*, Executive Office of the President (2022), <https://www.whitehouse.gov/wp-content/uploads/2022/02/CEQ-CEJST-QandA.pdf> [perma.cc/TCX3-PXZP]
- 82 Yuxin Xiao, Shulammit Lim, Tom Joseph Pollard, and Marzyeh Ghassemi, *In the Name of Fairness: Assessing the Bias in Clinical Record De-identification*, Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency (FAcCT '23), <https://doi.org/10.1145/3593013.3593982>. [perma.cc/X4RW-2LQH]

- 83 Joy Buolamwini and Timnit Gebru, *Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification*, Conference on Fairness, Accountability, and Transparency (2018), <https://proceedings.mlr.press/v81/buolamwini18a/buolamwini18a.pdf>. [perma.cc/HBC9-5EKK]
- 84 Luca Belli, Kyra Yee, Uthaipon Tantipongpipat, Aaron Gonzales, Kristian Lum and Moritz Hardt, *County-level Algorithmic Audit of Racial Bias in Twitter's Home Timeline*, arXiv (2013), <https://doi.org/10.48550/arXiv.2211.08667> [https://perma.cc/5DQT-JXWZ]; Zhaowei Zhu, Yuanshun Yao, Jiankai Sun, Hang Li, and Yang Liu, *Weak Proxies are Sufficient and Preferable for Fairness with Missing Sensitive Attributes*, Proceedings of the 40th International Conference on Machine Learning (2023), <https://proceedings.mlr.press/v202/zhu23n/zhu23n.pdf>. [https://perma.cc/LT63-8EJX]
- 85 Roy L. Austin Jr., "Race Data Measurement and Meta's Commitment to Fair and Inclusive Products," Meta Newsroom (November 2021), <https://about.fb.com/news/2021/11/inclusive-products-through-race-data-measurement>. [perma.cc/MR3R-77E3]
- 86 Kang-Xing Jin, "Supporting Equitable Access to COVID-19 Vaccines," Meta Newsroom (April 2021), <https://about.fb.com/news/2021/04/supporting-equitable-access-to-covid-19-vaccines>. [perma.cc/LSS4-47W8]
- 87 *United States v. Meta Platforms, Inc., f/k/a Facebook, Inc.* (2022) <https://www.justice.gov/crt/case-document/file/1514111/dl?inline>. [perma.cc/MK2S-PUUH]
- 88 Javier Martell, Paul Panichelli, Rich Strauch, and Sally Taylor-Shoff, *The Effectiveness of Scoring on Low-to-Moderate-Income and High-Minority Area Populations*, Fair Isaac (1991).
- 89 Jennifer O'Hara, "Socioeconomic status measure helps researchers develop artificial intelligence models, improving equity in health care," Mayo Clinic (December 2022), <https://newsnetwork.mayoclinic.org/discussion/socioeconomic-status-measure-helps-researchers-develop-artificial-intelligence-models-improving-equity-in-health-care> [perma.cc/S8EC-HW32]
- 90 Rishabh Mehrotra, James McInerney, Hugues Bouchard, Mounia Lalmas, and Fernando Diaz, *Towards a Fair Marketplace: Counterfactual Evaluation of the trade-off between Relevance, Fairness & Satisfaction in Recommendation Systems*, CIKM '18: Proceedings of the 27th ACM International Conference on Information and Knowledge, <https://doi.org/10.1145/3269206.3272027>. [perma.cc/ZD6Q-PWLS]
- 91 Drew Harwell, "The Accent Gap," The Washington Post (July 2018), <https://www.washingtonpost.com/graphics/2018/business/alexa-does-not-understand-your-accent>. [perma.cc/8GXC-5JC5]
- 92 Maya Gupta, Andrew Cotter, Mahdi Milani Fard, and Serena Wang, *Proxy Fairness*, arXiv (June 2018), <https://doi.org/10.48550/arXiv.1806.11212>. [perma.cc/QSM7-3NHL]

- 93 See, e.g. Andrew Myers, *AI-Detectors Biased Against Non-Native English Writers*, Stanford University Human-Centered Artificial Intelligence (2023), <https://hai.stanford.edu/news/ai-detectors-biased-against-non-native-english-writers>. [perma.cc/36MD-3NHQ]
- 94 For example, within the US government, agencies hold fragmented slices of data about citizens and beneficiaries, and while data sharing across federal agencies has increased, inconsistent policies and practices, legal barriers, and practical challenges mean that data sharing between agencies may be experienced with government as more akin to dealing with auxiliary datasets. *Data Sharing Working Group*, Federal CDO Council, <https://www.cdo.gov/data-sharing>. [https://perma.cc/B2RK-443A]
- 95 The US Government Accountability Office has requested that Congress revise laws to facilitate secure data sharing in order to facilitate disparity analysis. *Tax Equity: Lack of Data Limits Ability to Analyze Effects of Tax Policies on Households by Demographic Characteristics*, US Government Accountability Office (2022), <https://www.gao.gov/products/gao-22-104553>. [perma.cc/8DAV-XPTG]
- 96 *A Vision for Equitable Data: Recommendations from the Equitable Data Working Group*; Wally Adeyemo and Lily Batchelder, "Advancing Equity Analysis in Tax Policy," US Department of Treasury (December 2021), <https://home.treasury.gov/news/featured-stories/advancing-equity-analysis-in-tax-policy>. [perma.cc/23FJ-TBN2]
- 97 Hadi Elzayn, Evelyn Smith, and Thomas Hertz et al, *Measuring and Mitigating Racial Disparities in Tax Audits*, Stanford Institute for Economic Policy Research (2023), https://dho.stanford.edu/wp-content/uploads/IRS_Disparities.pdf. [perma.cc/EZ9F-8HF4]
- 98 Muhammad Ali et al, *Discrimination through Optimization: How Facebook's Ad Delivery Can Lead to Biased Outcomes*, Proceedings of ACM Human-Computer Interaction (CSCW), <https://doi.org/10.1145/3359301> [https://perma.cc/NF7Q-7JLD]; Jinyan Zang, *Solving the problem of racially discriminatory advertising on Facebook*, Brookings (2021), <https://www.brookings.edu/articles/solving-the-problem-of-racially-discriminatory-advertising-on-facebook>. [https://perma.cc/9CTM-LDE8]
- 99 Jeni Soucie, Bryan O. Buckley, Karri Albanese, Rachel Harrington, and Sarah Hudson Scholle, *Current Health Plan Approaches to Race and Ethnicity Data: Collection and Recommendations for Future Improvements*, National Committee for Quality Assurance (NCQA) (2023), https://www.ncqa.org/wp-content/uploads/2023/03/Current-Health-Plan-Approaches-to-Race-and-Ethnicity-Data-Collection-and-Recommendations-for-Future-Improvements_Final.pdf [https://perma.cc/7728-GUPY]; the approach CMS used to collect data is described in Celia Eicheldinger and Arthur Bonito, *More Accurate Racial and Ethnic Codes for Medicare Administrative Data*, Health Care Finance Review, <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4195038>. [https://perma.cc/SJF8-YJQ7]

- 100 In some cases, patients may be offered the opportunity to consent to data sharing, though the practice of asking for specific consent does not appear to be consistent. *Current Health Plan Approaches to Race and Ethnicity Data: Collection and Recommendations for Future Improvements*.
- 101 *Collection of Race and Ethnicity Data for Use by Health Plans to Advance Health Equity*, Urban Institute (2022), https://www.urban.org/sites/default/files/2022-07/Collection%20of%20Race%20and%20Ethnicity%20Data%20for%20Use%20by%20Health%20Plans%20to%20Advance%20Health%20Equity_final.pdf [perma.cc/6BH8-4YCN]
- 102 Denise M. Hynes, Matthew L. Maciejewski, and David Atkins, *HSR Commentary: Linking VA and Non-VA Data to Address Important US Veteran Health Services Research Issues*, Health Serv Res (2018), <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6235822>. [perma.cc/4CQ2-4KGR]
- 103 Elsa Augustine, Vikash Reddy, and Jesse Rothstein, *Linking Administrative Data: Strategies and Methods*, California Policy Lab (2018), <https://www.capolycylab.org/wp-content/uploads/2018/12/Data-Linking-white-paper-12-18-18-FINAL.pdf>. [perma.cc/VZ4V-YTJK]
- 104 James C Doidge and Katie L Harron, *Reflections on modern methods: linkage error bias*, International Journal of Epidemiology, <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7020770> [https://perma.cc/S9N6-SGAV]; Ruth Gilbert, Rosemary Lafferty, Gareth Hagger-Johnson, Katie Harron, Li-Chun Zhang, Peter Smith, Chris Dibben, and Harvey Goldstein, *GUILD: GUIDance for Information about Linking Data sets*, Journal of Public Health (2018), <https://doi.org/10.1093/pubmed/fox037>. [https://perma.cc/9VKD-BLVX]
- 105 Louise Mc Grath-Lone, Nicolás Libuy, David Etoori, and Ruth Blackburn, *Ethnic bias in data linkage*, *The Lancet* (2021), <https://www.thelancet.com/action/showPdf?pii=S2589-7500%2821%2900081-9> [https://perma.cc/65JL-QQ8A]; Joseph T. Lariscy, *Differential Record Linkage by Hispanic Ethnicity and Age in Linked Mortality Studies: Implications for the Epidemiologic Paradox*, Journal of Aging and Health, <https://doi.org/10.1177/0898264311421369>. [https://perma.cc/NHQ7-3BQ6]
- 106 Ridhi Shetty, "CDT Comments to CFPB Lay Out Data Broker Harms That Should Be Held Accountable," Center for Democracy & Technology (July 2023), <https://cdt.org/insights/cdt-comments-to-cfpb-lay-out-data-broker-harms-that-should-be-held-accountable>. [perma.cc/CLC5-YYV3]
- 107 Michael Veale and Reuben Binns, *Fairer machine learning in the real world: Mitigating discrimination without collecting sensitive data*, Big Data & Society (2017), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3060763 [perma.cc/QHK2-69AK]
- 108 Yusuf Talha Tamer, Albert Karam, Thomas Roderick, and Steve Miff, *Know Thy Patient: A Novel Approach and Method for Patient Segmentation and Clustering Using Machine Learning to Develop Holistic, Patient-Centered Programs and Treatment Plans*, NEJM Catalyst (August 23, 2022), <https://catalyst.nejm.org/doi/full/10.1056/CAT.22.0084>. [https://perma.cc/9ERV-PQFT]

- 109 Pranav Dheram, Murugesan Ramakrishnan, Anirudh Raju, I-Fan Chen, Brian King, Katherine Powell, Melissa Saboowala, Karan Shetty, and Andreas Stolcke, *Toward Fairness in Speech Recognition: Discovery and mitigation of performance disparities*, arXiv, <https://doi.org/10.48550/arXiv.2207.11345>. [perma.cc/7E8F-4ZWC]
- 110 David Liu, Virginie Do, Nicolas Usunier, and Maximilian Nickel, *Group fairness without demographics using social networks*, ACM Conference on Fairness, Accountability, and Transparency (FAccT '23), <https://doi.org/10.1145/3593013.3594091>. [perma.cc/CND9-GQDU]
- 111 Alexandra Chouldechova, Siqi Deng, Yongxin Wang, Wei Xia, and Pietro Perona, *Unsupervised and Semi-supervised Bias Benchmarking in Face Recognition*, Computer Vision – ECCV 2022 Lecture Notes in Computer Science, https://doi.org/10.1007/978-3-031-19778-9_17. [perma.cc/24ML-NPVG]
- 112 Pranav Dheram, Murugesan Ramakrishnan and Anirudh Raju et al, *Toward Fairness in Speech Recognition: Discovery and mitigation of performance disparities*, arXiv (2022), <https://doi.org/10.48550/arXiv.2207.11345>. [perma.cc/7E8F-4ZWC]
- 113 J. Khadijah Abdurahman, "FAT* Be Wilin," *Medium* (February 2019), <https://upfromthecracks.medium.com/fat-be-wilin-deb56bf92539>. [perma.cc/KF56-GYBK]
- 114 Julianna Rowsell, "Reducing biased and harmful outcomes in generative AI," Adobe Design (January 31, 2024) <https://adobe.design/stories/leading-design/reducing-biased-and-harmful-outcomes-in-generative-ai> [perma.cc/A5HZ-7D58]
- 115 Tolga Bolukbasi, Kai-Wei Chang, James Zou, Venkatesh Saligrama, and Adam Kalai, *Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings*, 30th Conference on Neural Information Processing Systems (NIPS 2016), <https://doi.org/10.48550/arXiv.1607.06520>. [perma.cc/2EVC-X4SU]
- 116 Emily Sheng, Kai-Wei Chang, Premkumar Natarajan, and Nanyun Peng, *The Woman Worked as a Babysitter: On Biases in Language Generation*, EMNLP 2019, <https://doi.org/10.48550/arXiv.1909.01326>. [perma.cc/SH47-VT53]
- 117 Ben Hutchinson, Vinodkumar Prabhakaran and Emily Denton, Kellie Webster, Yu Zhong, and Stephen Denuyl, *Social Biases in NLP Models as Barriers for Persons with Disabilities*, ACL (2020), <https://doi.org/10.48550/arXiv.2005.00813>. [perma.cc/6NW7-H8EM]
- 118 Adina Williams, Eric Smith, Rebecca Qian, and Melanie Kambadur, "Introducing two new datasets to help measure fairness and mitigate AI bias," Meta AI, <https://ai.meta.com/blog/measure-fairness-and-mitigate-ai-bias>.

- 119 Gabriel Nicholas and Aliyah Bhatia, *Lost in Translation: Large Language Models in Non-English Content Analysis*, Center for Democracy & Technology (2023), <https://cdt.org/insights/lost-in-translation-large-language-models-in-non-english-content-analysis> [<https://perma.cc/4PLH-DCQS>]; A. Stevie Bergman, Lisa Anne Hendricks, Maribeth Rauh, Boxi Wu, William Agnew, Markus Kunesch, Isabella Duan, Iason Gabriel, and William Isaac, *Representation in AI Evaluations*, ACM Conference on Fairness, Accountability, and Transparency (FAccT '23), <https://doi.org/10.1145/3593013.3594019>. [perma.cc/GGG8-7BAH]
- 120 Candice Schumann, Gbolahan O. Olanubi, Auriel Wright, Ellis Monk, Jr., Courtney Heldreth, and Susanna Ricco, *Consensus and Subjectivity of Skin Tone Annotation for ML Fairness*, 37th Conference on Neural Information Processing Systems (NeurIPS 2023), <https://doi.org/10.48550/arXiv.2305.09073> [perma.cc/PR7M-43KX] (describing that in addition to self-reported annotations, other methods of observed annotation include third party/crowdsourced annotation, annotations from captions, and algorithmic or model generated annotations).
- 121 Candice Schumann, Susanna Ricco, and Utsav Prabhu et al, *A Step Toward More Inclusive People Annotations for Fairness*, AIES '21, <https://dl.acm.org/doi/pdf/10.1145/3461702.3462594>. [perma.cc/865U-3MT7]
- 122 FACET Dataset, <https://ai.meta.com/datasets/facet>.
- 123 Victoria Turk, "How AI reduces the world to stereotypes," *Rest of World* (October 2023), <https://restofworld.org/2023/ai-image-stereotypes>. [perma.cc/6YND-HBMY]
- 124 "Building for Inclusivity: The Technical Blueprint of Pinterest's Multidimensional Diversification," Pinterest Engineering Blog (September 2023), <https://medium.com/pinterest-engineering/building-for-inclusivity-the-technical-blueprint-of-pinterests-multidimensional-diversification-a43d38840fb9>. [perma.cc/PS93-6KWB]
- 125 William Thong, Przemyslaw Joniak, Alice Xiang, *Beyond Skin Tone: A Multidimensional Measure of Apparent Skin Color*, International Conference on Computer Vision (ICCV) (2023), <https://ai.sony/publications/Beyond-Skin-Tone-A-Multidimensional-Measure-of-Apparent-Skin-Color>. [perma.cc/Q8UZ-HYNH]
- 126 Malaika Handa, "Colorism in High Fashion," *The Pudding* (2019), <https://pudding.cool/2019/04/vogue>. [perma.cc/ULN7-84ZZ]
- 127 "Meta Quest v57 Update: Unsend Image Messages, Horizon Feed, Avatars Update, and More," Meta Quest Blog (September 2023), <https://www.meta.com/blog/quest/v57-software-unsend-image-messages-horizon-home-avatars>.

- 128 Isabel Hexel, "Work In The Metaverse: Diversity And Liability Risks Employers Need To Consider," *Modaq* (September 2023), <https://www.mondaq.com/germany/discrimination-disability--sexual-harassment/1344210/work-in-the-metaverse-diversity-and-liability-risks-employers-need-to-consider>; Samantha Bielen, Wim Marneffe, and Naci Mocan, *Racial Bias and In-Group Bias in Virtual Reality Courtrooms*, *The Journal of Law and Economics* (May 2021), <https://www.journals.uchicago.edu/doi/abs/10.1086/712421>. [<https://perma.cc/PJ9N-EJPQ>]
- 129 E.g. Levi Kaplan, Nicole Gerzon, Alan Mislove, and Piotr Sapiezynski, *Measurement and Analysis of Implied Identity in Ad Delivery Optimization*, Proceedings of the 22nd ACM Internet Measurement Conference (IMC '22), <https://doi.org/10.1145/3517745.3561450> [perma.cc/44UU-27SP] ("[W]e refer to demographic information hinted at in the synthetic pictures as "implied" demographics. We make this distinction to avoid conflating self-reported demographic information of real individuals with stereotype-driven pixel perturbations").
- 130 Julienne LaChance, William Thong, Shruti Nagpal, and Alice Xiang, *A Case Study in Fairness Evaluation: Current Limitations and Challenges for Human Pose Estimation*, AAAI 2023 Workshop on Representation Learning for Responsible Human-Centric AI (R2HCAI), <https://r2hcai.github.io/AAAI-23/files/CameraReadys/21.pdf>. [perma.cc/2P6L-FF5C]
- 131 Ellis Monk, *The Monk Skin Tone Scale (MST)*, SocArXiv (2023), <https://doi.org/10.31235/osf.io/pdf4c> [<https://perma.cc/4SLN-W3WK>]
- 132 Trivellore E. Raghunathan, *Synthetic Data*, *Annual Review of Statistics and Its Application* (2021), <https://doi.org/10.1146/annurev-statistics-040720-031848>. [perma.cc/56JE-HGLE]
- 133 See e.g. Ivona Krchova, Michael Platzer, and Paul Tiwald, *Strong statistical parity through fair synthetic data*, NeurIPS 2023 Workshop on Synthetic Data Generation with Generative AI (2023), <https://doi.org/10.48550/arXiv.2311.03000>. [perma.cc/A6C3-SBVE]
- 134 See e.g. Wei-Yin Loh, Luxi Cao, and Peigen Zhou, *Subgroup identification for precision medicine: A comparative review of 13 methods*, *WIREs Data Mining and Knowledge Discovery* (2019), <https://wires.onlinelibrary.wiley.com/doi/abs/10.1002/widm.1326> [<https://perma.cc/ZD69-L36V>] (used simulated data to compare the performance of subgroup identification for precision medicine) and Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, and Krishna P. Gummadi, *Fairness Beyond Disparate Treatment & Disparate Impact: Learning Classification without Disparate Mistreatment*, WWW '17: Proceedings of the 26th International Conference on World Wide Web (2017), <https://doi.org/10.1145/3038912.3052660> [<https://perma.cc/6HWT-LTUD>] (uses synthetic datasets alongside real world datasets to demonstrate the potential efficacy of a fairness mitigation method); Alessandro Castelnovo, Riccardo Crupi, Nicole Inverardi, Daniele Regoli, and Andrea Cosentini, *Investigating Bias with a Synthetic Data Generator: Empirical Evidence and Philosophical Interpretation*, Proceedings

- of 1st Workshop on Bias, Ethical AI, Explainability and the Role of Logic and Logic Programming (BEWARE 2022), <https://doi.org/10.48550/arXiv.2209.05889> [<https://perma.cc/3G5N-RNK3>] (developing a synthetic data generator and testing it on mitigated and non-mitigated ML models); Sarah-Jane Van Els, David Graus, Emma Beauxis-Aussalet, *Improving Fairness Assessments with Synthetic Data: a Practical Use Case with a Recommender System for Human Resources*, Compjobs '22: the First International Workshop on Computational Jobs Marketplace (2022), https://compjobs.github.io/assets/paper_6.pdf [<https://perma.cc/3Z6E-QNEY>] (exploring the potential of synthetic data for assessing recommender system fairness, but finding there may be challenges in statistical reliability of using synthetic data for fairness assessment).
- 135 Alexander D'Amour, Hansa Srinivasan, James Atwood, Pallavi Baljekar, D. Sculley, and Yoni Halpern, *Fairness Is Not Static: Deeper Understanding of Long Term Fairness via Simulation Studies*, Conference on Fairness, Accountability, and Transparency (FAT* '20), <https://doi.org/10.1145/3351095.3372878>. [perma.cc/UE7U-5F5Q]
- 136 Pieter-Paul Verhaeghe, *Correspondence Studies*, Handbook of Labor, Human Resources and Population Economics (2022), https://doi.org/10.1007/978-3-319-57365-6_306-1 [<https://perma.cc/UX3H-JBQV>]; "Fair Housing Testing Program," US Department of Justice Civil Rights Division, <https://www.justice.gov/crt/fair-housing-testing-program-1> [<https://perma.cc/U3W3-52JF>] (accessed March 1, 2024); "Testing as a Civil Rights Mechanism to Prove Discrimination," NAACP (2023), <https://naacp.org/resources/testing-civil-rights-mechanism-prove-discrimination>. [<https://perma.cc/B7KG-ACTV>]
- 137 Marianne Bertrand and Sendhil Mullainathan, *Are Emily and Greg More Employable than Lakisha and Jamal? A Field Experiment on Labor Market Discrimination*, NBER Working Paper Series (2003), <http://www.nber.org/papers/w9873>. [perma.cc/F2LV-JUMN]
- 138 Naroa Martínez, Aranzazu Vinas, and Helena Matute, *Examining potential gender bias in automated-job alerts in the Spanish market*, PLOS One (2021), <https://doi.org/10.1371/journal.pone.0260409>. [perma.cc/PRU3-9P5D]
- 139 *Measurement and Analysis of Implied Identity in Ad Delivery Optimization*
- 140 Alex Tamkin, Amanda Askell, Liane Lovitt, Esin Durmus, Nicholas Joseph, Shauna Kravec, Karina Nguyen, Jared Kaplan, and Deep Ganguli, *Evaluating and Mitigating Discrimination in Language Model Decisions*, arXiv (2024), <https://doi.org/10.48550/arXiv.2312.03689> [<https://perma.cc/BAK8-KP46>]; see also Abel Salinas, Parth Vipul Shah, Yuzhong Huang, Robert McCormack, and Fred Morstatter, *The Unequal Opportunities of Large Language Models: Revealing Demographic Bias through Job Recommendation*, EAAMO '23, <https://arxiv.org/pdf/2308.02053.pdf>. [<https://perma.cc/K4C8-8BB7>]
- 141 Eve Fleisig, Aubrie Amstutz, Chad Atalla, Su Lin Blodgett, Hal Daumé III, Alexandra

- Olteanu, Emily Sheng, Dan Vann, and Hanna Wallach, *FairPrism: Evaluating Fairness-Related Harms in Text Generation*, Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics, <https://aclanthology.org/2023.acl-long.343>. [perma.cc/3TPW-DSJC]
- 142 Karan Bhanot, Miao Qi, John S. Erickson, Isabelle Guyon, and Kristin P. Bennett, *The Problem of Fairness in Synthetic Healthcare Data*, Entropy (2021), <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8468495>. [perma.cc/AY5P-VZSY]
- 143 *Improving Fairness Assessments with Synthetic Data: a Practical Use Case with a Recommender System for Human Resources*
- 144 *FairPrism: Evaluating Fairness-Related Harms in Text Generation*
- 145 *Fairer machine learning in the real world: Mitigating discrimination without collecting sensitive data*
- 146 *Human Rights Due Diligence of Meta's Impacts in Israel and Palestine in May 2021*, BSR (2022), https://www.bsr.org/reports/BSR_Meta_Human_Rights_Israel_Palestine_English.pdf [https://perma.cc/K4A7-WTNP]; see also Ryan Mac, "Instagram Censored Posts About One Of Islam's Holiest Mosques, Drawing Employee Ire," BuzzFeed News (May 2021), <https://www.buzzfeednews.com/article/ryanmac/instagram-facebook-censored-al-aqsa-mosque>. [https://perma.cc/K76Q-NLFJ]
- 147 Dave Gershgorn, "Companies are on the hook if their hiring algorithms are biased," Quartz (2018) <https://qz.com/1427621/companies-are-on-the-hook-if-their-hiring-algorithms-are-biased>. [https://perma.cc/M2H7-TUCL]
- 148 *Explainability & Fairness in Machine Learning for Credit Underwriting*, FinRegLab (2023), https://finreglab.org/wp-content/uploads/2023/12/FinRegLab_2023-12-07_Research-Report_Explainability-and-Fairness-in-Machine-Learning-for-Credit-Underwriting_Policy-Analysis.pdf [perma.cc/R2TW-GNJZ]
- 149 Tatsunori B. Hashimoto, Megha Srivastava, Hongseok Namkoong, and Percy Liang, *Fairness Without Demographics in Repeated Loss Minimization*, Proceedings of the 35th International Conference on Machine Learning (2018), <https://doi.org/10.48550/arXiv.1806.08010>. [perma.cc/29D9-T69T]
- 150 *Fairness for Unobserved Characteristics: Insights from Technological Impacts on Queer Communities*
- 151 *Year 1 Report. National Artificial Intelligence Advisory Committee (NAIAC).*
- 152 Anders Blok and Morten Axel Pedersen, *Complementary social science? Qualitative experiments in a Big Data world*, Big Data & Society (2014), <https://doi.org/10.1177/2053951714543908>. [https://perma.cc/HJG6-GM5G]
- 153 Angèle Christin, *The ethnographer and the algorithm: beyond the black box*, Theory

- and Society, Springer (2020), <https://doi.org/10.1007/s11186-020-09411-3> [<https://perma.cc/A6ZV-GM3A>]; Vidushi Marda and Shivangi Narayan, *On the importance of ethnographic methods in AI research*, Nature Machine Intelligence (2021), <https://doi.org/10.1038/s42256-021-00323-0>. [<http://perma.cc/68WE-U6T6>]
- 154 [Blueprint for an AI Bill of Rights](#)
- 155 Aaron Rieke, Miranda Bogen and David G. Robinson, *Public scrutiny of automated decisions: early lessons and emerging methods*, Upturn (2018), <https://apo.org.au/node/210086>. [<perma.cc/9FJK-9AS2>]
- 156 Chelsea Barabas, Colin Doyle, JB Rubinovitz, and Karthik Dinakar, *Studying Up: Reorienting the study of algorithmic fairness around issues of power*, ACM Conference on Fairness, Accountability, and Transparency, <https://doi.org/10.1145/3351095.3372859>. [<perma.cc/8HR6-K99U>]
- 157 Oliver L. Haimson and Anna Lauren Hoffmann, *Constructing and enforcing 'authentic' identity online: Facebook, real names, and non-normative identities*, First Monday, <https://doi.org/10.5210/fm.v21i6.6791>. [<perma.cc/VQ5V-4MJK>]
- 158 Kate Ruane, "Biden Must Halt Face Recognition Technology to Advance Racial Equity," ACLU (February 2021), <https://www.aclu.org/news/privacy-technology/biden-must-halt-face-recognition-technology-to-advance-racial-equity>. [<perma.cc/CJ8P-V3XU>]
- 159 [On the importance of ethnographic methods in AI research](#)
- 160 Stephanie Willson and Kristen Miller, *Cognitive Interview Evaluation of X Gender Marker Definitions for the U.S. Passport Application Form*, Collaborating Center for Question Design and Evaluation Research, National Center for Health Statistics (2022), https://wwwn.cdc.gov/QBank/report/Miller_2022_NCHS_GenderX.pdf [<https://perma.cc/C2H5-F5UC>]; *A Vision for Equitable Data: Recommendations from the Equitable Data Working Group*, White House (2022), <https://www.whitehouse.gov/wp-content/uploads/2022/04/eo13985-vision-for-equitable-data.pdf>. [<https://perma.cc/FR7K-P6VF>]
- 161 "Introduction: A Practical Guide to Inclusive Research," Inclusive Research (May 2021), <https://medium.com/inclusive-research/introduction-a-practical-guide-to-inclusive-research-8a3c87375b0e> [<perma.cc/7RG5-TZPU>]
- 162 "Product Equity at Adobe," Adobe, <https://adobe.design/product-equity-at-adobe>

- [<https://perma.cc/P355-Q5W6>] (accessed March 1, 2024); "Building for everyone, with everyone," Google, <https://about.google/belonging/product-inclusion-and-equity> [<https://perma.cc/F8HV-P852>] (accessed March 1, 2024); "Weaving equity into the way the world moves," Uber Blog (June 2022), <https://www.uber.com/blog/weaving-equity-into-the-way-the-world-moves/> [<https://perma.cc/C4M9-JH4P>]; Raena Saddler, Denasia Pinkard, Anne Diaz, Tulsee Doshi, Imani Dunbar, and Zach Singleton, "From Ideas to Impact: Insights from the December 2022 Product Equity Summit," Medium (February 2023), <https://medium.com/@techproductequity/from-ideas-to-impact-insights-from-the-december-2022-product-equity-summit-f08139c892d4>. [<https://perma.cc/D5YW-Q2JR>]
- 163 Timothy Bardlavens, "Charting the course for Adobe's Product Equity team," Adobe Design (May 2023), <https://adobe.design/stories/design-for-scale/charting-the-course-for-adobe-s-product-equity-team>. [perma.cc/W8NJ-33XP]
- 164 Molly McHugh-Johnson, "How we tested Guided Frame and Real Tone on Pixel," Google Blog (February 2024), <https://blog.google/products/pixel/how-we-tested-the-pixels-new-inclusive-camera-features>. [perma.cc/2CRP-PAKB]
- 165 *Fairness for Unobserved Characteristics: Insights from Technological Impacts on Queer Communities*
- 166 See e.g. "Representation Matters: A Recruiting Process for Inclusive UX Research," AnswerLab (April 2021), <https://www.youtube.com/watch?v=sRJUyXfQbKc>. [perma.cc/7KPX-ZK7A]
- 167 *Who Do They Think You Are? Categories, Classification, and Profiling*
- 168 *Collection of Race and Ethnicity Data for Use by Health Plans to Advance Health Equity*
- 169 *Review into bias in algorithmic decision-making*, Centre for Data Ethics and Innovation (2020), https://assets.publishing.service.gov.uk/media/60142096d3bf7f70ba377b20/Review_into_bias_in_algorithmic_decision-making.pdf. [perma.cc/WX8C-N95Z]
- 170 *Measuring discrepancies in Airbnb guest acceptance rates using anonymized demographic data*, Airbnb (2020), <https://news.airbnb.com/wp-content/uploads/sites/4/2020/06/Project-Lighthouse-Airbnb-2020-06-12.pdf>. [<https://perma.cc/5UE8-3CXF>]
- 171 *How Meta is working to assess fairness in relation to race in the U.S. across its products and systems*
- 172 "Algorithmic Discrimination," Blueprint for an AI Bill of Rights
- 173 "What is RBAC (Role-Based Access Control) – And is it Right For You?" Immuta, <https://www.immuta.com/guides/data-security-101/rbac-role-based-access-control> (accessed March 1, 2024). [perma.cc/83KL-7WNB]
- 174 Shizra Sultan and Christian D. Jensen, *Ensuring Purpose Limitation in Large-Scale*

- Infrastructures with Provenance-Enabled Access Control*, Sensors (2021) <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8123646> [perma.cc/3L85-ZQZ2]; “About Limited Data Use,” Meta Business Help Center, <https://www.facebook.com/business/help/1151133471911882> (accessed March 1, 2024); “Data Processing Options for US Users,” Meta for Developers, <https://developers.facebook.com/docs/marketing-apis/data-processing-options> (accessed March 1, 2024).
- 175 [Measuring discrepancies in Airbnb guest acceptance rates using anonymized demographic data](#)
- 176 Joaquin Quiñonero-Candela, Yuwen Wu, Brian Hsu, Sakshi Jain, Jen Ramos, Jon Adams, Robert Hallman, and Kinjal Basu, *Disentangling and Operationalizing AI Fairness at LinkedIn*, ACM Conference on Fairness, Accountability, and Transparency (FAccT '23), <https://doi.org/10.1145/3593013.3594075>. [perma.cc/4S5Y-UFU4]
- 177 [How Meta is working to assess fairness in relation to race in the U.S. across its products and systems](#)
- 178 “Attribute-Based Access Control,” Immuta, <https://www.immuta.com/product/secure/attribute-based-access-control> [https://perma.cc/GH9Z-AFNL] (accessed March 1, 2024); *NIST Researchers Publish Book on Attribute-Based Access Control*, NIST Computer Security Resource Center (CSRC) (2018), <https://csrc.nist.gov/News/2018/NIST-Researchers-Publish-Book-on-ABAC> [https://perma.cc/4JHP-4AN8]; [Ensuring Purpose Limitation in Large-Scale Infrastructures with Provenance-Enabled Access Control.](#)
- 179 [Explainability & Fairness in Machine Learning for Credit Underwriting](#)
- 180 *See Unlocking the value of data: Exploring the role of data intermediaries*, Centre for Data Ethics and Innovation (2021), <https://www.gov.uk/government/publications/unlocking-the-value-of-data-exploring-the-role-of-data-intermediaries>. [perma.cc/3V5T-3L4N]
- 181 [Report: Enabling responsible access to demographic data to make AI systems fairer](#)
- 182 [Fairer machine learning in the real world: Mitigating discrimination without collecting sensitive data](#)
- 183 [How Meta is working to assess fairness in relation to race in the U.S. across its products and systems](#)
- 184 Keith Bonawitz, Hubert Eichner and Wolfgang Grieskamp et al, *Towards Federated Learning at Scale: System Design*, Proceedings of the 2nd SysML Conference (2019), <https://arxiv.org/pdf/1902.01046.pdf> [https://perma.cc/HGL3-FP3Q]; Daniel Ramage and Stefano Mazzocchi, “Federated Analytics: Collaborative Data Science without Data Collection,” Google Research Blog (May 2020), <https://blog.research.google/2020/05/federated-analytics-collaborative-data.html?m=1>. [https://perma.cc/EH35-NVTN]
- 185 For a detailed discussion of federated statistics as a potential approach to facilitate

- fairness measurement, see Sarah Villeneuve, Tina M. Park and Eliza McCullough, *Eyes Off My Data: Exploring Differentially Private Federated Statistics To Support Algorithmic Bias Assessments Across Demographic Groups*, Partnership on AI (023), https://partnershiponai.org/wp-content/uploads/dlm_uploads/2023/12/PAI_whitepaper_eyes-off-my-data-1.pdf [https://perma.cc/6HPG-GW9K]
- 186 *Eyes Off My Data*
- 187 "What is encryption?" Cloudflare Learning Center, (accessed March 1, 2024), <https://www.cloudflare.com/learning/ssl/what-is-encryption>.
- 188 Airbnb further describes the asymmetric encryption process: "Using a public key provided by Research Partner, File 1 is asymmetrically encrypted prior to persisting so that, because only Research Partner has the private key, Airbnb can no longer decrypt File 1 (which is why it lies within the Research Partner Trust Boundary) and thus loses the linking from nid to photo_url." *Measuring discrepancies in Airbnb guest acceptance rates using anonymized demographic data*.
- 189 E.g. Paulo Martins, Leonel Sousa, and Artur Mariano, *A Survey on Fully Homomorphic Encryption: An Engineering Perspective*, ACM Computer Surveys (2018), <https://doi.org/10.1145/3124441> [perma.cc/5JED-3XUB]; Louis J. M. Aslett, Pedro M. Esperanca and Chris C. Holmes, *A review of homomorphic encryption and software tools for encrypted statistical machine learning*, arXiv, <https://doi.org/10.48550/arXiv.1508.06574>. [perma.cc/5BE9-AMKA]
- 190 In some cases, longer retention periods are required. For instance, educational institutions must retain race and ethnicity data for 3 years (or longer in the event of litigation, audit, or similar actions). *Collecting Race And Ethnicity Data from Students and Staff Using the New Categories*, National Center for Education Statistics, <https://nces.ed.gov/ipeds/report-your-data/race-ethnicity-collecting-data-for-reporting-purposes> [perma.cc/GWV6-9EN4]
- 191 *Eyes Off My Data*
- 192 *Measuring discrepancies in Airbnb guest acceptance rates using anonymized demographic data; How Meta is working to assess fairness in relation to race in the U.S. across its products and systems*
- 193 "Welcoming platform demographic survey," Nextdoor Help Center, https://help.nextdoor.com/s/article/Welcoming-platform-demographic-survey?language=en_US [https://perma.cc/J83M-JZZA] (accessed March 1, 2024); "How LinkedIn uses your personal demographic data," LinkedIn Help (March 2024), <https://www.linkedin.com/help/linkedin/answer/a1336543>. [https://perma.cc/A3AU-SY72]
- 194 *How Meta is working to assess fairness in relation to race in the U.S. across its products and systems*

- 195 See e.g. Khaled El Emam and Fida Kamal Dankar, Protecting Privacy Using k-Anonymity, *Journal of the American Medical Informatics Association* (2008), <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2528029>. [perma.cc/55M9-MVE9]
- 196 *Measuring discrepancies in Airbnb guest acceptance rates using anonymized demographic data*
- 197 *How Meta is working to assess fairness in relation to race in the U.S. across its products and systems*
- 198 *Eyes Off My Data*
- 199 Alessandro Fabris, Andrea Esuli, Alejandro Moreo, and Fabrizio Sebastiani, *Measuring Fairness Under Unawareness of Sensitive Attributes: A Quantification-Based Approach*, *Journal of Artificial Intelligence Research* (2023), <https://www.jair.org/index.php/jair/article/view/14033/26912>. [perma.cc/NAG6-F44Q]
- 200 Cynthia Dwork and Aaron Roth, *The Algorithmic Foundations of Differential Privacy*, *Foundations and Trends in Theoretical Computer Science* (2014), <https://www.cis.upenn.edu/~aaroht/Papers/privacybook.pdf>. [perma.cc/K9BU-Y2XQ]
- 201 Amalie Dyda, Michael Purcell, and Stephanie Curtis et al, *Differential privacy for public health data: An innovative tool to optimize information sharing while protecting data confidentiality*, *Patterns* (2021), <https://doi.org/10.1016/j.patter.2021.100366>. [perma.cc/Z777-S6BA]
- 202 Miguel Guevara, "How we're helping developers with differential privacy," *Google for Developers* (January 2021), <https://developers.googleblog.com/2021/01/how-were-helping-developers-with-differential-privacy.html> [https://perma.cc/8DD6-UT5V]; Chaya Nayak, "New privacy-protected Facebook data for independent research on social media's impact on democracy," *Meta Research* (February 2020), <https://research.facebook.com/blog/2020/2/new-privacy-protected-facebook-data-for-independent-research-on-social-medias-impact-on-democracy>; "Apple Differential Privacy Technical Overview," *Apple*, https://www.apple.com/privacy/docs/Differential_Privacy_Overview.pdf [https://perma.cc/8TZ3-ETHE].
- 203 The Census adopted differential privacy in particular for the 2020 census. "Statistical Safeguards," *United States Census Bureau*, https://www.census.gov/about/policies/privacy/statistical_safeguards.html [https://perma.cc/E3DS-XSSX]. Their deployment generated controversy. Priyanka Nanayakkara and Jessica Hullman, "States Are Suing the Census Bureau Over Its Attempts to Make Data More Private," *Slate* (August 2021), <https://slate.com/technology/2021/08/census-bureau-differential-privacy-lawsuit.html> [https://perma.cc/9PKZ-JLYQ].
- 204 For a detailed discussion of Apple's efforts, see *Eyes Off My Data*.
- 205 Miranda Bogen, Pushkar Tripathi, and Aditya Srinivas Timmaraju et al, *Towards Fairness in Personalized Ads*, *Meta* (2023), https://about.fb.com/wp-content/uploads/2023/01/Toward_fairness_in_personalized_ads.pdf. [perma.cc/Q49B-Q7B6]

- 206 For a discussion of local DP see Hussein Mozannar, Mesrob I. Ohannessian, and Nathan Srebro, *Fair Learning with Private Demographic Data*, arXiv (2020), <https://doi.org/10.48550/arXiv.2002.11651> [<https://perma.cc/YD6A-GG26>]; see also Marc Juarez and Aleksandra Korolova, 'You Can't Fix What You Can't Measure': Privately Measuring Demographic Performance Disparities in Federated Learning, Proceedings of Machine Learning Research (2023), <https://arxiv.org/pdf/2206.12183.pdf>. [<https://perma.cc/DJK3-4VBB>]
- 207 According to the Partnership on AI, an epsilon value ≤ 1 is considered maximally private, a value between 2 and 10 is considered to provide some privacy, and a value > 10 is considered to provide little to no privacy, though this depends on context. *Eyes Off My Data*.
- 208 The ϵ across different contexts may need to differ to accomplish the same privacy goals depending on the size of a dataset, since larger datasets may be able to tolerate more statistical noise while retaining accuracy than smaller datasets can. *Guidelines for Evaluating Differential Privacy Guarantees*.
- 209 *Towards Fairness in Personalized Ads*
- 210 Ryan Steed et al, *Policy impacts of statistical uncertainty and privacy*, Science (2022), <https://www.science.org/doi/10.1126/science.abq4481>.
- 211 *Do No Harm Guide: Applying Equity Awareness in Data Privacy Methods*, Urban Institute (2023), <https://www.urban.org/research/publication/do-no-harm-guide-applying-equity-awareness-data-privacy-methods>. [perma.cc/44HZ-7WMN]
- 212 Niki Kilbertus, Adria Gascon and Matt Kusner et al, *Blind Justice: Fairness with Encrypted Sensitive Attributes*, Proceedings of the 35th International Conference on Machine Learning (2018), <https://arxiv.org/pdf/1806.03281.pdf>. [perma.cc/Q88Z-UXYW]
- 213 "Understanding How Different Communities Experience Instagram," Instagram Blog (July 2022), <https://about.instagram.com/blog/announcements/collecting-and-measuring-demographic-information>; for a detailed description of Meta's implementation of secure multi-party computation, see *How Meta is working to assess fairness in relation to race in the U.S. across its products and systems*.
- 214 Elizabeth Renieris, "Why PETs (privacy-enhancing technologies) may not always be our friends," Ada Lovelace Institute (April 2021), <https://www.adalovelaceinstitute.org/blog/privacy-enhancing-technologies-not-always-our-friends>. [perma.cc/N7ZX-UJUD]
- 215 "How LinkedIn uses your personal demographic data," <https://www.linkedin.com/help/linkedin/answer/a1336543> (accessed March 1, 2024). [perma.cc/U2YF-Y7W8]
- 216 "A new way we're fighting discrimination on Airbnb"
- 217 Facebook Privacy Policy, <https://www.facebook.com/privacy/policy> (accessed March 11, 2024)
- 218 *Eyes Off My Data*

- 219 [EU AI Act](#)
- 220 [How Meta is working to assess fairness in relation to race in the U.S. across its products and systems](#)
- 221 [Eyes Off My Data](#)
- 222 [Explainability & Fairness in Machine Learning for Credit Underwriting](#)
- 223 [Measuring discrepancies in Airbnb guest acceptance rates using anonymized demographic data](#)
- 224 On the other hand, relative understaffing of fairness-relevant teams can constrain their capacity to proactively detect issues. Hayden Field, “How Airbnb failed its own anti-discrimination team—and let racial disparities slip through the cracks,” Tech Brew (June 2021), <https://www.emergingtechbrew.com/stories/2021/06/15/airbnb-failed-antidiscrimination-team-and-let-racial-disparities-slip-cracks>. [perma.cc/SA3W-SQTN]
- 225 FinRegLab has described that “[s]maller firms may have difficulty attracting and maintaining equal levels of expertise to their modeling and compliance teams. As a result, model developers and statisticians from business units may provide limited support for certain compliance activities and may have limited access to protected class information in the course of implementing less discriminatory alternatives.” [Explainability & Fairness in Machine Learning for Credit Underwriting](#).
- 226 [Explainability & Fairness in Machine Learning for Credit Underwriting](#)
- 227 “Privacy Impact Assessments,” Electronic Privacy Information Center, <https://epic.org/issues/open-government/privacy-impact-assessments> (accessed March 1, 2024). [perma.cc/XJ8L-SYDU]
- 228 [Eyes Off My Data; Do No Harm Guide: Applying Equity Awareness in Data Privacy Methods](#)
- 229 [Eyes Off My Data](#)
- 230 “COC Partnership Helps Airbnb Formulate Anti-Discrimination Plans,” Color of Change (February 2023), <https://colorofchange.org/coc-partnership-helps-airbnb-formulate-anti-discrimination-plans> [perma.cc/U9Y7-JSAR]
- 231 [Eyes Off My Data](#)
- 232 [Do No Harm Guide: Applying Equity Awareness in Data Privacy Methods](#)
- 233 Elisa Johnson, “Aiming for truth, fairness, and equity in your company’s use of AI,” Federal Trade Commission Business Blog (April 2021), <https://www.ftc.gov/business-guidance/blog/2021/04/aiming-truth-fairness-equity-your-companys-use-ai>. [perma.cc/MAC3-9ZPY]
- 234 [Using publicly available information to proxy for unidentified race and ethnicity: A methodology and assessment](#)

- 235 “RE: Comprehensive Use of Civil Rights Authorities to Prevent and Combat Algorithmic Discrimination,” Upturn, Lawyer’s Committee for Civil Rights Under Law, and the Center for Democracy & Technology (Feb 2024), https://cdt.org/wp-content/uploads/2024/02/Letter_to_DOJ_re_AI_Executive_Order.pdf. [<https://perma.cc/G4LH-ENY9>]
- 236 *Advisory Committee on Data for Evidence Building: Year 2 Report*, Advisory Committee on Data for Evidence Building (2022), <https://www.bea.gov/system/files/2022-10/acdeb-year-2-report.pdf>.
- 237 *Exploring legal mechanisms for data stewardship*, Ada Lovelace Institute (2021), https://www.adalovelaceinstitute.org/wp-content/uploads/2021/03/Legal-mechanisms-for-data-stewardship_report_Ada_AI-Council-2.pdf. [perma.cc/LXC8-H5SG]
- 238 *Who Do They Think You Are? Categories, Classification, and Profiling*


 cdt.org


 cdt.org/contact

 **Center for Democracy & Technology**

1401 K Street NW, Suite 200

Washington, D.C. 20005

 202-637-9800

 @CenDemTech

