

Participatory and Inclusive Demographic Data Guidelines

Eliza McCullough
Sarah Villeneuve

Use in conjunction with:

- [Implementation Workbook](#)
- [Case Study](#)



**Version for Public Comment
Not for Distribution**

Contents

Executive Summary	3
Introduction	4
Why Did We Create the Guidelines?	4
What Are the Goals of the Guidelines?	6
Who Are These Guidelines for?	7
Participatory and Inclusive Demographic Data Guidelines	8
Guidelines Across the Demographic Data Lifecycle	13
Overview	13
1. Planning & Design	14
2. Consent	16
3. Collection	18
4. Pre-Processing	20
5. Analysis	22
6. Bias Mitigation & Results-Sharing	24
7. Removal, Archive, & Destruction	26
8. Documentation & Data Governance	27
Acknowledgments	28
Appendix 1: Glossary	29

Version
for Public
Comment

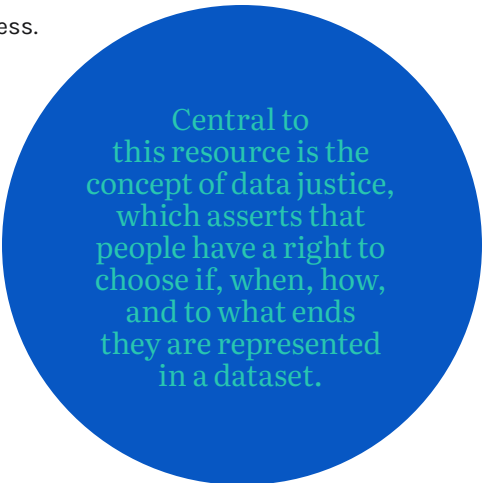
Executive Summary

Too often, the **algorithmic systems*** used to determine outcomes in various settings discriminate against historically **marginalized groups**. In response, many organizations that develop or use algorithmic systems turn to fairness assessments to measure the presence of statistical bias, which requires collecting and analyzing **demographic data**.

* Phrases highlighted in yellow are defined in the Glossary starting on page 29.

However, while demographic data is often necessary to identify and rectify algorithmic discrimination, Partnership on AI's [previous research](#) revealed that incautious collection and use of this data could lead to a secondary set of substantial harms against data subjects, particularly those from marginalized groups. These harms present a key tension: the apparent need to collect demographic data to address algorithmic discrimination and the imperative to prevent harms that can stem from this very process.

The Participatory and Inclusive Demographic Data Guidelines sit at the intersection of this dilemma. The Guidelines aim to provide AI developers, teams within technology companies, and other data practitioners with **guidance on collecting and using demographic data for fairness assessments to advance the needs of data subjects and communities, particularly those most at risk of harm from algorithmic systems**. Central to this resource is the concept of data justice, which asserts that people have a right to choose if, when, how, and to what ends they are represented in a dataset.



Central to this resource is the concept of data justice, which asserts that people have a right to choose if, when, how, and to what ends they are represented in a dataset.

In collaboration with our Working Group and broader multistakeholder community, we developed actionable guidelines for organizations based on the demographic data lifecycle (see [Figure 3](#)). We include *key risks* faced by data subjects and communities (particularly marginalized groups) during each lifecycle phase, *recommended practices* that organizations should undertake to prevent these risks, and *guiding questions* for organizations and practitioners to reflect on as they implement the practices. We also identify four *key principles* across the demographic data lifecycle:

- Prioritize the Right to Self-Identification
- Co-Define Fairness
- Implement Affirmative, Informed, Accessible, & Ongoing Consent
- Promote Equity-Based Analysis

This resource aims to promote fairness measurement efforts that benefit marginalized communities. We envision a world where everyone, including those from socially marginalized groups, can utilize and benefit from algorithmic systems.

Introduction

Why Did We Create the Guidelines?

Algorithmic systems are increasingly used to make decisions in a variety of settings.

Too often, these systems discriminate against historically marginalized groups such as LGBTQIA+ people (particularly gender non-conforming and transgender individuals), Black and Indigenous people, racially and ethnically marginalized people, people with disabilities, and people from the Majority World.* Organizations that develop and/or use algorithmic systems often turn to quantitative fairness testing using demographic data to detect and address discriminatory outcomes.

For example, a company might build an algorithmic system that screens resumes and flags the strongest ones. The company may want to assess whether the system is operating fairly across demographic groups for various reasons, ranging from the desire to ensure equitable product performance to regulatory compliance. To do so, they may collect relevant demographic information from people impacted by the algorithmic system, in this case, job applicants who submit their resumes for review. The company could then use that data to determine whether the system flagged resumes from members of certain demographic groups at disproportionate rates. If the system flagged resumes from white men at a higher rate than resumes from other applicants, they could use these results to inform a fairness intervention, which could include adjusting system components, retraining the system, further investigating other sources of bias, or discarding the system altogether.

But while demographic data is often necessary to identify and rectify algorithmic discrimination, we know that the collection and use of this data can lead to substantial harms against individuals and communities (particularly those from historically marginalized groups) from [previous PAI research](#). These harms can include the expansion of surveillance infrastructure, the misrepresentation of the use of sensitive data beyond data subjects' expectations, and the misrepresentation of what it means to hold a certain identity (see [Figure 1](#)). These negative results are possible even when companies collect and use demographic data with the intent to mitigate harms from biased decision-making.

In our previous example, the company's demographic data collection process could survey applicants' gender but only provide binary gender options, which could erase any trans or gender non-conforming applicants and possible biases these applicants face. The company could also wrongfully repurpose the demographic dataset collected for the fairness assessment to inform job ad-targeting efforts even though participants only consented to using their dataset for the defined assessment. Finally, the company could fail to securely store the dataset and expose sensitive information to data leaks (such as data related to sexuality) that put applicants at risk of harm or violence. These are just some of the

* The term "Majority World" was proposed by scholars from what was formerly referred to as the "Third World." As Dr. Shahidul Alam explains, "The term highlights the fact that we are indeed the majority of humankind and brings to sharp attention the anomaly that the Group of eight countries—whose decisions affect the majority of the world's peoples—represent a tiny fraction of humankind." ([Source](#))

possible negative outcomes of data collection with insufficient guardrails.

These harms present a key tension: the apparent need to collect demographic data to address algorithmic discrimination and the imperative to prevent the harms that can stem from this very process. Often, the communities most at risk of algorithmic discrimination are also [most vulnerable](#) to the harms caused by demographic data collected to rectify this discrimination. **The Guidelines aim to address this tension and promote fairness measurement efforts that benefit socially marginalized communities. We envision a world in which all people, including those from socially marginalized groups, can utilize and benefit from algorithmic systems.**

FIG. 1 Risks Associated with Demographic Data Collection

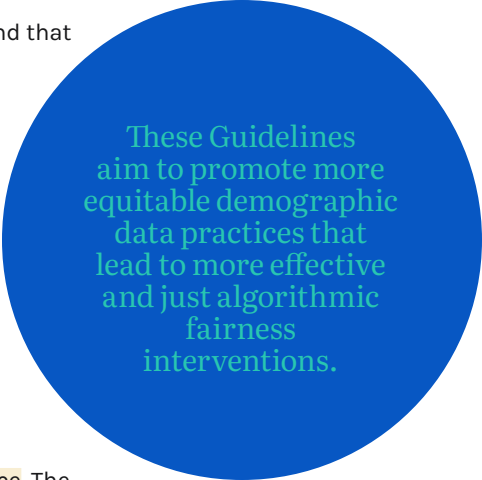
RISKS & HARMS TO INDIVIDUALS	DEFINITION
Privacy risks due to leaks of sensitive identity information	As attributes such as race, ethnicity, country of birth, gender, religion, and sexuality are usually consequential aspects of one’s identity, collecting and using this data presents key privacy risks. Usage of demographic information can allow for harmful consequences such as political ad targeting of marginalized groups, leading to racial inequities in information access.
Miscategorization and misrepresentation of individuals in the data collection process	Individual misrepresentation can lead to discrimination and disparate impacts. Algorithmically inferred racial category collection practices can further entrench pseudoscientific practices that assume invisible aspects of one’s identity from visible characteristics, such as physiognomy .
Use of sensitive data beyond data subjects’ expectations	The use of demographic information beyond its initial intent not only violates the consent of data subjects but can also lead to unintended harmful consequences. For example, the US government developed the Prisoner Assessment Tool Targeting Estimated Risk and Needs to provide guidance on recidivism reduction programming, but the tool was later repurposed to inform inmate transfers which led to racially disparate outcomes.
RISKS & HARMS TO COMMUNITIES	
Expansion of surveillance infrastructure in the name of fairness	Marginalized communities are often subjected to invasive, cumbersome, and experimental data collection methods, which fairness assessments can further exacerbate. Expanded surveillance can constrain these communities’ agency and result in their exploitation. Data collected from marginalized communities is often used against them, such as in predictive policing technology and other law enforcement surveillance tactics.
Misrepresenting and mischaracterizing what it means to be part of a demographic group or to hold a certain identity	Incorrectly assigned demographic categories can reinforce harmful stereotypes, naturalize schemas of categorization, and cause other forms of “administrative violence.” This can occur because the range of demographic categories is too narrow, such as leaving out options for “non-binary” or “gender-fluid” in the case of gender, which leads to the undercounting of gender non-conforming individuals. Similarly, the available demographic categories may leave out key components of identity, such as the exclusion of disability status, which leads to the undercounting of people with disabilities .
Data subjects ceding the ability to define for themselves what constitutes biased or unfair treatment	When companies leading the data collection effort alone define fairness without input from marginalized groups, they can miss key instances of discrimination and reinforce the status quo. Strictly formalized definitions of fairness measurement can lead to ineffective and even harmful fairness interventions because they ignore the socio-historical conditions that lead to inequities.

What Are the Goals of the Guidelines?

These Guidelines aim to provide AI developers, teams within technology companies, and other **data practitioners** with **guidance on how to collect and use demographic data for statistical fairness assessment to advance the needs of data subjects and communities, particularly those most at risk of harm from algorithmic systems**. The Guidelines are informed by the **data justice** framework, which asserts that people have a right to choose if, when, how, under what circumstances, and to what ends they are represented in a dataset. These Guidelines aim to promote more equitable demographic data practices that lead to more effective and just algorithmic fairness interventions.

We recognize that statistical analysis alone is insufficient to capture the range of harms that algorithmic systems can cause. At various points throughout the Guidelines, we call on teams to consider alternative information sources when investigating bias, such as interviews, focus groups, and ethnographic studies. We also recommend that teams integrate perspectives from **sociotechnical** experts throughout the demographic data lifecycle, as sociotechnical analysis helps bring to light various social components influencing the design, production, and use of algorithmic systems and their social impacts. However, most of the Guidelines are tailored toward teams conducting **statistical bias measurements**. We chose this focus as it is the most commonly used approach to fairness assessments by AI-developing and AI-using organizations.

We also recognize that historically, the algorithmic fairness framework has **failed** to take into account a broader analysis of **algorithmic injustice**. The controversy over ProPublica's **investigation** of the COMPAS recidivism algorithm in 2016, which found that the algorithm achieves **predictive parity** across racial groups but still exacerbated racial inequities in incarceration by classifying Black defendants as higher risk than white defendants, highlights this discrepancy. An algorithmic system can meet statistical definitions of fairness but still reproduce and magnify existing social inequities. While we attempt to broaden the algorithmic fairness framework at key points, such as in Guiding Principle #2, we recognize the limitations of the 'fairness' framework. Note that this resource focuses on using demographic data to **support** fairness assessments rather than fairness measurement techniques.¹ Our Guidelines should not be used as a substitute for a broader analysis of harms that can be enabled by algorithmic systems or **critical engagement** with business practices that result in these harms. Instead, our Guidelines should be used in conjunction with other equity-promoting actions, advocacy work, and regulatory efforts that aim to advance algorithmic justice.



These Guidelines aim to promote more equitable demographic data practices that lead to more effective and just algorithmic fairness interventions.

¹ Miranda Bogen, "Navigating Demographic Measurement for Fairness and Equity," Center for Democracy & Technology, forthcoming May 2024.

FUTURE POLICY DIRECTIONS

As these Guidelines are targeted toward AI-developing and AI-using organizations (in both the public and private sectors), we focused on actionable recommendations to change existing demographic data practices. However, we recognize that policy is crucial in advancing the equitable use of algorithmic systems. Achieving data justice for all requires that people enjoy comprehensive data rights that protect their agency over the collection and use of their digital identities. We encourage future collaboration on designing a robust policy agenda to further advance participatory and inclusive demographic data practices for algorithmic fairness.

Who Are These Guidelines for?

Our primary audience for these Guidelines are **organizations that develop and/or use AI and are tasked with conducting an algorithmic fairness assessment that involves demographic data collection**. We hope these Guidelines also serve as a resource for civil society and impacted communities engaged in algorithmic justice advocacy, as well as public agencies, vendors, and other third-party stakeholders involved in the procurement, use, and assessment of algorithmic systems. The Guidelines are also intended to inform policymakers' efforts to build policy governing algorithmic fairness and demographic data collection and usage.

This resource was developed specifically for demographic data collected **in pursuit of fairness interventions**, keeping in mind the specific tradeoffs and risks faced by data subjects for this use case. Collecting information about identity is inherently risky for data subjects and communities. However, the threshold for allowable risk may be higher when an explicit benefit, such as the promise of a fairer algorithmic system, is a potential outcome for such data collection and analysis.

We recognize that organizations may collect demographic data for various purposes beyond fairness assessments, like user research and ad-targeting. Many tech business models also rely on the collection and processing of user behavioral data as a proxy for demographic information. Aspects of these Guidelines may be useful to the companies collecting demographic information for goals beyond fairness (namely our Guiding Principles 1, 3, and 4). However, we encourage companies using demographic data for other purposes to think critically about the risks to data subjects highlighted in this resource.

Finally, these Guidelines are best suited for demographic data collection to assess bias in static algorithmic systems, which are designed to perform a certain task or output, rather than dynamic AI models which are designed to continuously adapt and can be re-trained post-deployment. In the future, we hope to adapt these Guidelines to address the specific discrimination challenges posed by dynamic models.

HOW DID WE CREATE THESE GUIDELINES?

Partnership on AI (PAI) began our drafting process with a literature review of relevant themes, including [data governance](#), [data equity](#), and [data justice](#). We then convened the Participatory and Inclusive Demographic Data Working Group, composed of members from the technology industry, academia, civil society, and government offices such as Markkula Center for Applied Ethics, Data Economy Policy Hub, DeepMind, Apple, and Center for Democracy & Technology based in the US, UK, Canada, South Africa, the Netherlands, and Australia. The Working Group met monthly from January 2023 to March 2024 to discuss and co-draft each component of the Guidelines. We also gathered feedback from attendees at workshops hosted at [Mozilla Festival](#), [Data Justice Conference](#), and Partnership on AI's [2023 Partner Forum](#), [Columbia School of Social Work](#), [National Housing Conference Racial Equity Working Group](#), [Natural Sciences and Engineering Research Council of Canada](#), and PAI [Demographic Data Quarterly Community Meetings](#).

PAI is based in the US, and the conferences where we presented the Guidelines all took place in the Global North. While some members of our Working Group are from or reside in the Majority World, most work and live in Europe or the US. Given the limitations of our perspectives, we commissioned seven equity experts residing in the Majority World who specialize in data justice to review the Guidelines. We also commissioned a disability rights expert, a racial justice expert, and a data justice expert to provide an in-depth review given the relevance of their expertise to the Guidelines. Note that the current version of the Participatory and Inclusive Demographic Data Guidelines is the result of a participatory, iterative multistakeholder process and should not be read as representing views from individual contributors or organizations.

Participatory and Inclusive Demographic Data Guidelines

GUIDING PRINCIPLES

The following reflect the four key principles that inform all of our Guidelines across the demographic data lifecycle. We include a *brief description* and references to *relevant lifecycle phases* for each principle.

1. Prioritize the Right to Self-Identification

Demographic data collection methods can be thought of on a continuum, in which methods that allow for no self-identification (like [inference techniques](#)) exist at one end while methods that allow for full self-identification (like write-in surveys) exist on the other. Methods such as multiple-choice surveys exist somewhere in the middle. Organizations often opt for inference techniques and [methods](#) with minimal self-identification as they can pose fewer privacy risks and generate more efficient and robust datasets. Data collection methods that allow for more self-identification require significantly more time and resourcing and can also lead to smaller datasets, limiting certain forms of analysis (including [potential privacy loss](#)). However, as our previous research demonstrates,

RELEVANT
LIFECYCLE PHASES

1. Planning & Design
2. Consent
3. Collection
4. Pre-Processing
5. Analysis

failure to prioritize self-identification in demographic data collection can result in a range of harms for data subjects and communities, from the reinforcement of oppressive stereotypes to the obfuscation of key algorithmic harms.

Organizations should make every effort to provide data subjects and communities with agency over how their identities are represented when collecting demographic data for fairness assessments. While the ability to fully capture socially constructed demographic identities in a dataset is [inherently limited](#), organizations should strive to minimize the gap between how data subjects understand themselves and how they are represented in the dataset. We urge organizations to prioritize data collection methods that allow for data subjects to self-describe or self-select their demographic traits, work directly with data subjects to identify relevant demographic categories, and conduct ongoing checks with data subjects to ensure that the representation of their identities is valid and accurate. These ongoing checks should still adhere to the principle of [data minimization](#) by ensuring that only necessary information is collected and held for the minimum amount of time possible.

DEMOGRAPHIC DATA COLLECTION & INFERENCE METHODS

These Guidelines focus primarily on self-identified demographic data (see Guiding Principle 1) in which data subjects can choose or describe their own demographic traits rather than statistical techniques that infer demographics through proxy information (like [image classification models](#) that assess peoples' skin tone or [Bayesian Improved Surname and Geocoding](#)) or [synthetic data](#) generation in which an artificial demographic dataset is created to mimic user information. Although this focus aligns with the concept of [data justice](#), we recognize that inference methods are often more accessible to companies assessing their models for bias. Inference methods can be used on existing datasets, such as user selfies or addresses, which cuts down on collection and processing costs, time, and organizational buy-in. It can also ensure sufficient coverage where fairness work may be critical or required but otherwise impossible. Similarly, synthetic data can be used to overcome some privacy risks. Since achieving high participation rates in direct, self-identified data collection can be difficult, inference methods applied to existing datasets can also potentially result in more robust fairness analyses (by achieving a larger sample size). Many advocates have called attention to these methods' [limitations](#) and potential to erase, misidentify, and obscure discrimination. Companies that pursue inference methods should still follow these Guidelines and draw on additional resources to ensure that their use of inference methods is just, equitable, and inclusive.²

2 Lockhart, Jeffrey W., Molly M. King, and Christin L. Munsch. "Computer Algorithms Infer Gender, Race and Ethnicity. Here's How to Avoid Their Pitfalls." *Nature*, July 5, 2023. <https://doi.org/10.1038/d41586-023-02225-0>; Randall, Megan. "Five Ethical Risks to Consider before Filling Missing Race and Ethnicity Data," n.d. ; Rieke, Aaron, Vincent Southerland, Dan Svirsky, and Mingwei Hsu. "Imperfect Inferences: A Practical Assessment." In *2022 ACM Conference on Fairness, Accountability, and Transparency*, 767-77. Seoul Republic of Korea: ACM, 2022. <https://doi.org/10.1145/3531146.3533140>. ; Wojcik, Stefan. "The Challenges of Using Machine Learning to Identify Gender in Images." Pew Research Center (blog), September 5, 2019. <https://www.pewresearch.org/internet/2019/09/05/the-challenges-of-using-machine-learning-to-identify-gender-in-images/>.

2. Co-Define Fairness

Many sociotechnical [experts have noted](#) that fairness is an “essentially contested concept,” meaning it has “multiple context-dependent, and sometimes even conflicting, theoretical understandings.”³ While an organization may choose to define a ‘fair’ algorithmic system as one that is equally accurate for users across demographic groups, the users of the system may define ‘fairness’ as a system that benefits – or otherwise generates positive outcomes for – all users regardless of their demographic identities. Similarly, an organization may define an ‘unfair’ system as one that results in [disparate impacts](#) across identities, while users may define an ‘unfair’ system as one that [contributes to structural inequities](#). These discrepancies can lead to fairness assessments that satisfy the organization’s concerns but fail to meet the needs of (and potentially harm) those impacted by the algorithmic system, including those who provide their demographic data to support the fairness assessment. The process of defining ‘fairness’ for an algorithmic assessment is often limited to the statistical analysis team, with little to no involvement of those impacted by the system.

We urge organizations to work towards aligning their definition of ‘fairness’ with data subjects’ and communities’ expectations of ‘fairness’ when collecting demographic data for bias assessments. To achieve this, organizations and data subjects must also align on what constitutes ‘unfairness’ in the algorithmic system by agreeing on the types of harm or discrimination that warrant investigation through the fairness assessment.

These Guidelines are focused on demographic data to **support** fairness assessments rather than fairness measurement techniques.⁴ However, we recognize that alignment on what a ‘fair’ (and ‘unfair’) algorithmic system means is crucial for designing demographic data collection and usage practices that advance the needs of data subjects and communities. Organizations should work with data subjects and communities in co-defining ‘fairness’ to conduct data collection practices that advance the needs of this population (see [Figure 2](#)). Organizations should also make the operating definitions of ‘unfairness’ and ‘fairness’ used in the fairness assessment process publicly available to allow for adequate transparency to regulators and advocates and drive best practices.

RELEVANT
LIFECYCLE PHASES

- 1. Planning & Design
- 5. Analysis
- 6. Bias Mitigation & Results-Sharing

3 Corbett-Davies, Sam, Johann D. Gaebler, Hamed Nilforoshan, Ravi Shroff, and Sharad Goel. “The Measure and Mismeasure of Fairness.” arXiv, August 14, 2023. <https://doi.org/10.48550/arXiv.1808.00023>.

4 Miranda Bogen, “Navigating Demographic Measurement for Fairness and Equity,” Center for Democracy & Technology, forthcoming May 2024

3. Implement Affirmative, Informed, Accessible, & Ongoing Consent

Consent practices for data collection are often lacking in terms of accessibility, clarity, and functionality. In the US, for example, organizations are [not required](#) to expressly seek consent when collecting user data under most jurisdictions. Strong consent practices are key to advancing data justice by giving data subjects agency over how, when, under what conditions, and to what ends they are represented digitally. Organizations must prioritize and seek time-bound, opt-in consent. While organizations should create an explicit opportunity for data subjects to opt-in to participate in the Consent phase, they should also continue to consider participants' ongoing consent throughout the lifecycle as frequently as possible (such as by providing participants with the ability to revoke consent or re-submit consent if the implications of participation change). The consent process should be [clear, approachable, and accessible](#) to data subjects, particularly those most at risk of harm by the algorithmic system.

Prioritizing affirmative, informed, ongoing, and accessible consent can require considerable organizational resources and result in smaller sample sizes, which can prevent certain forms of analysis. We also recognize that revocation of consent can be limited in the context of algorithmic systems and urge organizations to communicate these limitations to data subjects at the very start of data collection.

4. Promote Equity-Based Analysis

Statistical bias assessments often employ techniques that assess whether the algorithmic system works for most users. This is reflected in measurements such as the [80 percent rule](#), a popular fairness metric that tests for instances of adverse impact by calculating whether the selection rate for a minority group (the group with the lowest selection rate) is less than 80 percent of the rate for the group with the highest selection rate (typically the majority group). While useful for identifying large discrepancies in selection rates, the 80 percent rule [can fail](#) to adequately uncover algorithmic harms experienced by statistical minorities, often users from marginalized demographic groups.

We urge organizations to focus on the needs and risks of groups most at risk of harm by the algorithmic system throughout the demographic data lifecycle. This involves a variety of practices such as consultation with members of or advocates for these communities to understand their needs during the Planning & Design phase or specific outreach to these communities during the Collection phase. Demographic categories, as well as power dynamics between demographic groups, are highly context-specific. Organizations must engage with data subjects and communities impacted by their algorithmic system to understand which groups are most at risk of harm. Embracing an equity-based analysis can ensure communities most at risk of algorithmic harm benefit from the demographic data collection process and create [cascading benefits](#) for all data subjects and user groups.

RELEVANT LIFECYCLE PHASES

- 2. Consent
- 3. Collection
- 5. Analysis
- 7. Removal, Archive, & Destruction
- 8. Documentation & Data Governance

RELEVANT LIFECYCLE PHASES

- 1. Planning & Design
- 2. Consent
- 4. Pre-Processing
- 5. Analysis
- 6. Bias Mitigation & Results-Sharing

FIG. 2 Stakeholder Engagement Methods

Well-designed, participatory, stakeholder engagement methods are key to preventing technology-mediated harms, including algorithmic discrimination, by allowing impacted communities to provide direct input into AI design and use. They involve a continuum of methods, ranging from data subject consultation to full decision-making empowerment.

DEGREE OF PARTICIPATION 

CONSULT	INVOLVE	COLLABORATE	EMPOWER
Data subjects chosen by the fairness assessment team give input on design via questionnaires, ranking decision alternatives, etc. The assessment process is fully determined by the fairness assessment team. Data subjects are likely not directly notified about the fairness assessment outcome.	The fairness assessment team chooses data subjects to participate in facilitated group discussions to gather input. The assessment process is primarily determined by the fairness assessment team but includes some room for data subjects' input. Data subjects may be notified about the fairness assessment outcome.	The fairness assessment team works directly with data subjects in ongoing collaborative design, prototyping, and decision-making about the fairness assessment process. The assessment process is co-determined by the fairness assessment team and data subjects. Data subjects are directly notified about the fairness assessment outcome.	Data subjects meaningfully contribute to key decisions about the fairness assessment process. The assessment process is determined by the data subjects with input from the fairness assessment team. Data subjects are directly notified about the fairness assessment outcomes on an ongoing basis.

Table adapted from "Stakeholder Participation in AI: Beyond 'Add Diverse Stakeholders and Stir'"

While crucial in advancing algorithmic justice, [researchers](#) and [advocates](#) have also called attention to how participatory methods can be used to manufacture community approval, resulting in "[participation washing](#)." We urge AI-developing and AI-using organizations, particularly those in the private sector, to pursue [participatory methods](#) in their fairness assessments while acknowledging that all participation is a form of labor that should be recognized, addressing inherent power asymmetries in stakeholder engagement, and integrating participatory methods across the demographic data lifecycle. In addition, organizations must be transparent about the goals of participatory processes to build and maintain trust with communities. PAI's [Global Task Force for Inclusive AI](#) will release a framework in the summer of 2024 that outlines practices that enable close coordination and partnership between AI developers, deployers, and the communities impacted by the technological change, as well as the guardrails that protect communities from increased risk and harm. For more information, sign up for our [mailing list](#).

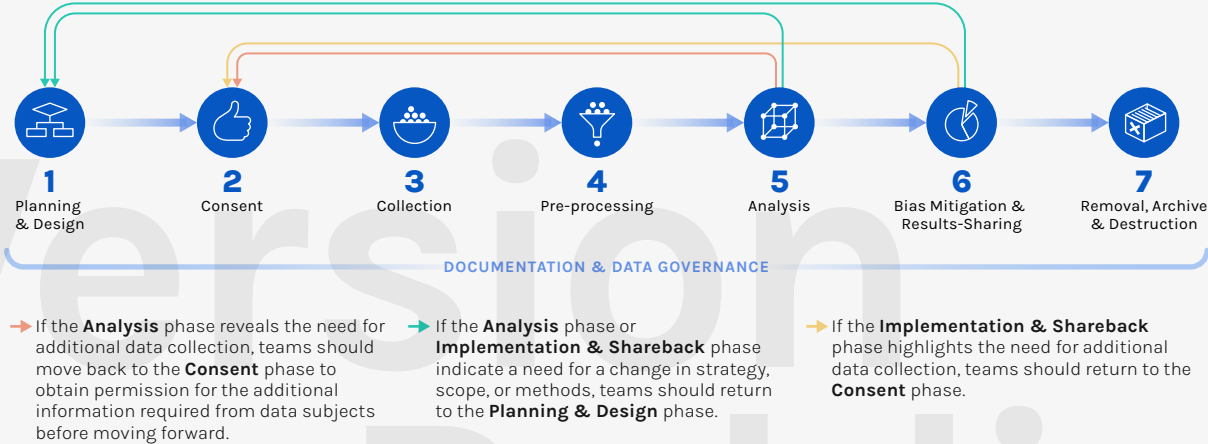
Version for Public Comment

Guidelines Across the Demographic Data Lifecycle

Overview

The Guidelines are organized by phases in the demographic data lifecycle. Each lifecycle phase includes a *description*. We then outline *key risks* faced by data subjects and communities (particularly **marginalized groups**) during the lifecycle phase and suggest practices that organizations should undertake to prevent these risks. These include *Baseline Requirements* (or the minimum practices that organizations should follow to mitigate risks against data subjects and communities) and *Recommended Practices* (or practices that organizations should follow to best **advance the needs** of data subjects and communities) for each lifecycle phase. Finally, we list *Guiding Questions* that organizations and practitioners should reflect on to mitigate risks faced by data subjects and communities and achieve the Recommended Practices outlined for the phase. We encourage organizations and teams to also use our companion Implementation Workbook for further guidance.

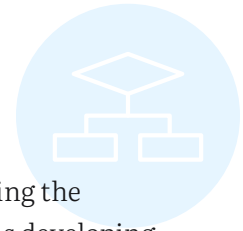
FIG. 3 The Demographic Data Lifecycle



The demographic data lifecycle describes the sequence of phases that a particular unit of demographic data goes through, from its initial collection to deletion or archival. As demographic data lifecycle phases can differ slightly across sources, we reviewed approximately 15 resources from industry, civil society, and academic publications. With consultation from our Working Group, we synthesized the Data Lifecycle into the following phases, which provided the structure for our Guidelines.

Version for Public Comment

1 Planning & Design



The planning and design phase of the demographic data lifecycle involves identifying the intended outcomes, relevant stakeholders, and viable collection methods, as well as developing a project plan for the collection and use of demographic data to support a fairness assessment.

KEY RISKS

- Impacted communities potentially harmed by the algorithmic system (directly or indirectly) are excluded from the Planning & Design process and do not get to inform or determine demographic categories collected or collection methods.
- Data practitioners lack adequate funding, time, or organizational buy-in, which leads to inadequate data collection and engagement from impacted communities.
- Power dynamics among stakeholders are ignored, leading to inequitable participation in the fairness assessment process.
- In the case of post-release system assessment, data practitioners do not consider the harms of allowing a potentially discriminatory system to remain in use when designing the fairness assessment process and timeline.

BASELINE REQUIREMENTS

- Team includes more than one **sociotechnical** expert in planning and design.
- Team indirectly integrates perspectives of impacted communities (i.e., through literature reviews and third-party conduits).
- Team has enough funding and other resources to complete the fairness assessment process but the resourcing is inflexible and cannot be adapted to unforeseen obstacles.
- Data practitioners practice **data minimization** when designing data collection and fairness assessment processes.

RECOMMENDED PRACTICES

- Experts from various disciplinary backgrounds, including those with lived experience relevant to the fairness assessment, are included in planning and design as paid members of the development team.
- Impacted communities potentially harmed by the algorithmic system are directly involved in designing the fairness assessment process, including identifying relevant demographic data.
- Data practitioners have sufficient funding and timeline to conduct a robust, inclusive, and participatory data collection and fairness assessment process.
- Data practitioners are trained to recognize and make efforts to address power imbalances between data subjects and data practitioners, as well as among data subjects across identity groups. They account for imbalances in process design through **tools** such as implicit bias training, recognizing positionality, and building structural competency.
- Organizational leadership is bought into designing and executing a robust and equitable demographic data and fairness assessment process.
- Teams provide data subjects and impacted communities with specific and clear mechanisms to contest the very existence of systems if they feel that they are fundamentally inequitable.

- Data practitioners take into account the harms of allowing a potentially discriminatory system to remain in use when designing the fairness assessment timeline and work to minimize it without sacrificing the quality of the assessment.

GUIDING QUESTIONS

- How will we define a successful outcome for the fairness assessment?
 - *What steps can we take to align this definition with that of other relevant parties (namely, data subjects and communities most at risk of harm)?*
- How will we obtain organizational commitment to attempt remediation if bias is identified in the fairness assessment?
 - *Will data subjects have the power to get the organization to stop using systems that they feel fundamentally cannot be made fair or be built in a way that does not harm them?*
 - *When a system is red-lit, will communities be able to access an equitable alternative?*
- What sociotechnical expertise (e.g., racial, gender, and other social equity experts) might be relevant to our fairness assessment?
 - *How can we ensure this expertise is incorporated at appropriate times?*
- What communities do we suspect are most at risk of harm by the algorithmic system (drawing on input from relevant equity experts and community leaders)?
 - *How will we prioritize them throughout the fairness assessment process?*
 - *How will we employ stakeholder engagement methods to include these communities?*
- What demographic attributes do we anticipate will be salient for this fairness assessment?
 - *How can we get feedback on this assumption from relevant community members and/or experts?*
- How will we practice data minimization in the fairness assessment process?
- Which factors may limit our ability to conduct a robust fairness assessment, including data availability for marginalized demographic categories? How will we overcome these constraints?
- Is our funding and timeline sufficient and flexible?

Version
for Public
Comment

2 Consent



After designing the data collection process and identifying relevant stakeholders, data subjects must be given an opportunity to agree or disagree to provide their data for a fairness assessment. Informed consent is relevant and may be required for multiple phases throughout the data lifecycle.

KEY RISKS

- Data subjects do not fully understand the implications of consent to participation in the data collection process, including limits to the right to data erasure and/or right to withdraw consent in the context of algorithmic systems and how they might benefit and/or be harmed through participation.
- Data architectures are insufficiently flexible to allow data subjects to consent to only some uses of their data and not others, forcing data subjects to make a binary decision to participate.
- Data subjects fear negative consequences (such as further exclusion or marginalization) if they do not consent.
- Data subjects are forced to give ‘blanket consent’ (i.e., their opt-in is not time-bound, specific, or revocable).
- Data subjects and communities are responsible for rejecting excessive or harmful data extraction.
- Data subjects are not provided multiple checkpoints to give or revoke consent.
- Members of demographic groups historically harmed by formal data collection processes consent at lower rates, leading to the exclusion of these groups from the dataset.
- Data practitioners are unable to fulfill promises about consent revocation.
- The consent process is opt-out rather than opt-in, meaning the data subjects’ consent is not affirmative.

BASELINE REQUIREMENTS

- Data subjects are informed about the types of data and measurements being collected, and for what purpose.
- Data subjects are given the ability to freely, actively, and affirmatively consent to or refuse participation.
- Data subjects can revoke consent at certain points in the fairness assessment process. Data practitioners are equipped to handle revocations by designing the system to allow the fairness assessment to continue in the event of an unexpected revocation. Additionally, data subjects can be removed from all future datasets, analyses, and training/testing processes.
- Data practitioners inform data subjects of any limitations to revocation of consent.
- Data practitioners inform data subjects about the implications of participation, communicating any serious risks.
- The language and format (i.e., written, oral, video, etc.) used for obtaining consent are [clear, approachable, and accessible](#) to data subjects, particularly those most at risk of harm by the algorithmic system.
- Data practitioners practice data minimization when requesting consent from data subjects and limit consent to only the use cases, data types, and time period necessary for the fairness assessment.

★ RECOMMENDED PRACTICES

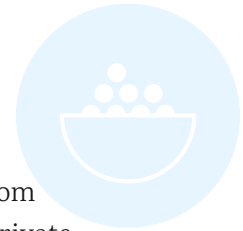
- Data practitioners work to build trust with leaders and representatives of communities historically harmed by formal data collection processes to ensure equitable representation in the dataset.
- Affected groups are consulted on the consent process, which is iteratively developed with feedback.
- Data subjects are given various tools (e.g., informational videos, articles, and virtual help desks) to support their understanding and agency during the consent process.
- Data practitioners design a consent revocation process that protects data subject and community-level privacy and is accessible to subjects at any point in the fairness assessment process.
- If members from demographic groups that have been historically harmed by formal data collection processes consent at lower rates, data practitioners conduct outreach to these communities and work to understand and correct issues in the consent process, leading to lower participation.

? GUIDING QUESTIONS

- How will we determine the unique needs of each group we are engaging?
 - *How can we ensure our consent process is accessible and clear to data subjects across demographic groups and not coercive or predatory?*
- How can we learn about the historical context of communities historically harmed by formal data collection efforts and use that context to inform our consent process?
- What is the expected life span of the dataset, and how will we inform data subjects of this expected life span?
- How can we operationalize the process of revoking consent and make it as easy as possible?
- How will we handle data in the case of consent revocation, and how will we protect the privacy of the data subject and communities in this process?
- Are there elements of participation that cannot be revoked?
 - *If so, how do we ensure data subjects are aware of these limits at the very start of the assessment process?*
- Can we provide data subjects with checkpoints throughout the data lifecycle that allow them to re-opt in or revoke consent (i.e., **dynamic consent**)?
- How do we ensure that data subjects are informed about the implications of their opt-in throughout the data lifecycle?

Version
for Public
Comment

3 Collection



Data collection can begin after consent is obtained. This involves gathering data from external sources through the method(s) of choice, such as public crowdsourcing, private sourcing, or automated data collection techniques.

KEY RISKS

- Impacted communities are oversampled in the collection process, leading to expanded surveillance of this group (see [Figure 1](#)).
- Impacted communities are undersampled in the collection process, leading to the invisibility of this group.
- Data subjects and impacted communities are not represented in the demographic dataset in how they define themselves (i.e., incorrect language, missing categories, etc), leading to individual or group-level **miscategorization**, reinforcement of harmful stereotypes, and invisibilization.
- Data subjects and impacted communities most at risk of harm are excluded from the data collection process due to accessibility barriers (e.g., lack of translation services, compatibility with assistive technology for people with disabilities, or alternative forms of outreach).
- Only quantitative datasets are collected, rather than surveys or interviews, thus obscuring certain (i.e., anecdotal or intersectional) instances of bias.
- Data subjects can only choose one or a limited amount of identities in each demographic category, leading to their misrepresentation in the dataset.
- Insufficient privacy-preserving techniques lead to demographic data **leakages**, **deanonymization**, and other privacy risks.
- Data is collected from data subjects who did not consent to the collection or who revoked their consent.

BASELINE REQUIREMENTS

- The collection process is accessible to the intended populations (e.g., interface, platforms, languages).
- The collection approach is validated by experts and data subjects (e.g., via pilots) and leads to a reasonably representative sample.
- Strong privacy-preserving techniques are implemented, reducing the risk of data leakages, deanonymization, and other risks.
- The data collection approach allows data subjects to express or select multiple identity groups.
- Data subjects and communities, particularly those most at risk of harm by the algorithmic system, can advise on the collection process (i.e., through an advisory group, community forum, or other feedback method).

Version for Public Comment

★ RECOMMENDED PRACTICES

- Data collected is representative of the user population, including impacted populations.
- Data practitioners augment structured demographic data collection mechanisms with open-ended forms of information collection, such as interviews and surveys.
- The data collection approach is rigorously validated (e.g., for selection bias and representativeness).
- Data practitioners work closely with data subjects to minimize the difference between how data subjects define themselves and how they are represented in the data.
- Data minimization is practiced throughout the data collection process.
- Privacy-preserving techniques used are designed to reflect the specific privacy concerns of the participant population, as privacy needs can differ depending on context.

? GUIDING QUESTIONS

- What is the minimum necessary information required to conduct our fairness assessment?
 - *How will we uphold the principle of data minimization as we collect data?*
- Can our chosen data collection method allow data subjects to identify themselves across multiple identities?
- If self-identification is not possible, how can we still minimize the inaccuracies between how data subjects define themselves and how they are represented in the data (such as by integrating feedback from interviews or focus groups with relevant stakeholders)?
- What are the open-ended methods to complement the data collection (i.e., interviews, focus groups, community forums)?
- How will we use an intersectional lens to define the representativeness of our collected dataset (i.e., checking for representation of identities across multiple, overlapping identities)?
- What are the limitations and benefits (particularly for communities most at risk of harm) of the available collection methods according to experts across multiple disciplines, community leaders, and data subjects?
- How will we validate whether the data collected through the chosen method is sufficient to satisfy the fairness assessment?

Version for Public Comment

4 Pre-Processing



Cleaning and preparing the data is necessary to make it usable for the fairness assessment. Data may be annotated, reformatted, summarized, or otherwise standardized in this phase.

KEY RISKS

- Specific harms against data subjects resulting from intersecting identities (i.e., someone who experiences both racial and gender discrimination) are lost during pre-processing.
- Pre-processing techniques (i.e., data cleaning and aggregation) degrade the quality of the data, particularly for groups that make up a **statistical minority** in the dataset.
- Insufficient privacy-preserving techniques lead to demographic data leakages, deanonymization, and other privacy risks.
- Privacy-preserving techniques disproportionately degrade the quality of data for minority groups.
- If data requires annotation, data annotators do not have appropriate working conditions, situated knowledge, or diversity of identities (across race, gender, ethnicity, nationality, disability, etc.), which impacts their ability to accurately annotate a demographically diverse dataset.

BASELINE REQUIREMENTS

- Data practitioners apply robust privacy-preserving techniques to reduce the risk of data leakages, deanonymization, and other privacy risks.
- Pre-processing techniques preserve the data quality by not recategorizing an individual into a single identity if they identify with multiple identities.
- If data requires annotation, data annotators have appropriate working conditions (including sufficient time and resources) and diverse identities (across race, gender, ethnicity, nationality, disability, etc.).
- Inference-based approaches to data annotation are only used for **carefully considered** characteristics as a last resort and in consultation with community advocates and/or equity experts.

RECOMMENDED PRACTICES

- Multiplicity of data subjects' identities is not diminished through data processing.
- Data is processed (i.e., cleaned, aggregated, or reformatted) so that any forms of discrimination that might have been captured in the data are clear and measured.
- Practitioners collect additional qualitative information from groups whose data cannot be analyzed without compromising privacy (due to the small sample size), such as via interviews or focus groups.

GUIDING QUESTIONS

- How will we approach typical data issues (e.g., missing data, label uncertainty, conflicting labels) without further erasing or harming marginalized communities?
- What privacy-enhancing approaches will we apply, and are they adequate to protect from human error or attacks?
- Do these approaches address the contextual privacy concerns of each marginalized community?

- If annotation is required, are annotator identities diverse and representative of those in the dataset?
 - *If not, what steps can be taken to ensure that annotator perspectives match the backgrounds necessary to assess the data?*
 - *Do annotators have healthy working conditions that allow them to prioritize quality over quantity of annotations?*
 - *Do annotators have the agency to provide input on assigned demographic categories, particularly if they have lived experiences relevant to people in the demographic categories?*
- How can we maintain intersectional representation of identities during pre-processing?
- What steps will we take if processing misrepresents certain demographic groups due to inadequate data?

5 Analysis

Next, this data is used to support executing a fairness assessment of the algorithmic system in question. The type of assessment depends on the operational definition(s) of [fairness](#), which can include demographic parity, predictive parity, or equalized odds, among others.

KEY RISKS

- Original goals of analysis (i.e., the goals that data subjects consented to when they opted in) are ignored.
- The fairness definition(s) used in the analysis (e.g., [demographic parity](#), [predictive parity](#), and [equalized odds](#)) obscure or do not effectively surface certain harms experienced by data subjects and impacted communities.
- The analysis process fails to look for particular harms occurring at the intersection of identities.
- Data practitioners do not consider how privacy-preserving techniques may obscure harms against demographic groups that make up a minority of the dataset and fail to choose a level of privacy that maintains data utility for these groups (while still protecting privacy).
- Impacted communities are not included in the analysis process (i.e., in design, discussion of results, etc.) to weigh in on the forms of discrimination or related harms that should be assessed.
- Only certain kinds of findings (i.e., those gained from quantitative analysis of algorithmic systems) are valued.
- Data practitioners fail to ensure that the collected dataset is analyzed in compliance with the informed consent obtained from data subjects.

BASELINE REQUIREMENTS

- Original goals of analysis (i.e., the goals that data subjects consented to when they opted in) are maintained.
- The analysis process is rigorous, comprehensive, well-documented, and reproducible.
- Data practitioners make every effort to ensure that bias against statistical minorities in the dataset is not obscured during the analysis.

★ RECOMMENDED PRACTICES

- Data practitioners employ an intersectional dataset analysis, assessing data subjects' potential experiences of discrimination across multiple and overlapping forms of demographic identity.
- Data practitioners supplement quantitative analysis of demographic datasets with alternative forms of analysis, such as findings from interviews and surveys.
- Data practitioners consult with data subjects and impacted communities, particularly those most at risk of harm by the algorithmic system, to ensure that the analysis addresses all relevant forms of discrimination and harm.

? GUIDING QUESTIONS

- What is our acting hypothesis about existing biases in the algorithmic system(s), and how will we test this hypothesis through the fairness analysis?
- How are we choosing our fairness analysis method, and what are the implications of these choices?
- Has the analysis involved only one type of analysis method (i.e., quantitative analysis), or have we used other necessary methods where relevant?
- How will stakeholders, particularly those from marginalized groups, be involved in the fairness analysis? What information about the analysis will be shared, with whom, and when?
- How will we engage experts from various disciplinary backgrounds in the fairness analysis?
- What potential edge cases might we have in our analysis?
- How can we address these?

6 Bias Mitigation & Results-Sharing

Once the fairness analysis is complete, an organization can identify the appropriate path to an appropriate fairness intervention, if necessary. The results of this fairness assessment and subsequent interventions should also be shared with relevant stakeholders.

! KEY RISKS

- Data practitioners fail to address the bias found during the analysis phase.
- Attempts to address the bias found during the analysis lead to new bias and harm against data subjects.
- The teams tasked with the assessment are not given sufficient power and funding to mitigate bias in the system.
- Findings are used to discount the lived experiences of impacted communities.
- Findings from the assessment are not made accessible to impacted communities.
- Data practitioners use fairness intervention as proof that all biases are eliminated even though data subjects and communities continue to experience harm.

✓ **BASELINE REQUIREMENTS**

- Data practitioners work to mitigate identified biases in the analysis.
- Data practitioners conduct a public results-sharing process that reaches data subjects and consulted communities, particularly those most at risk of harm by the algorithmic system.

★ **RECOMMENDED PRACTICES**

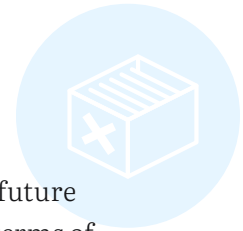
- Any findings of bias, discrimination, and related harms are effectively and thoughtfully addressed, leading to a more fair and just algorithmic system.
- Data practitioners red-light the algorithmic system if bias and discrimination cannot be mitigated.
- Data practitioners conduct a results-sharing process that involves direct, two-way discussion with data subjects and consulted communities about relevant fairness assessment findings.
- Data practitioners ensure the results-sharing process is accessible and provided in various formats.
- Data practitioners gather feedback from data subjects about their participation in the data collection, address feedback whenever possible, and conduct additional investigations when necessary.

? **GUIDING QUESTIONS**

- How do we plan to mitigate biases discovered in the fairness assessment with knowledge from impacted communities?
 - *How do we plan to share actions taken with data subjects?*
- How will we evaluate the success of fairness intervention(s) employed?
- What further steps must be taken if our actions do not adequately address any discrimination or harm found?
- How will we accessibly communicate the limitations or nuances of the fairness assessment (e.g., uncertainty, effect of privacy-enhancing approaches), particularly to groups found to have faced discrimination or harm?
- Have we provided adequate opportunities for data subjects and impacted communities to provide feedback?
 - *How will this feedback inform further or future investigations?*
- How will we make the shareback process accessible across demographic groups?

Version
for Public
Comment

7 Removal, Archive, & Destruction



Once the data has been used to support a fairness assessment, it may be stored for future use, removed from active environments, or destroyed securely, depending on the terms of consent obtained from the data subjects.

KEY RISKS

- Data is destroyed before all relevant analyses are completed, undermining measurement and remediation efforts and necessitating more or repetitive data collection.
- Data practitioners archive, re-use, share, or sell the dataset in ways inconsistent with the data subjects' consent.
- The data destruction process is ineffective or incomplete.
- Data practitioners fail to safely archive datasets, leading to data leakages, deanonymization, and other privacy harms.
- Data practitioners store data beyond the time limits of data subjects' consent.

BASELINE REQUIREMENTS

- Data practitioners safely destroy all copies of the dataset immediately after completion of its use consented to by data subjects.
- Data practitioners only archive datasets for future activities if the data subjects agree to them during the consent process.
- If data subjects consent to future use of the dataset, data practitioners safely archive the dataset and mitigate data leakages, deanonymization, and other privacy harms.

RECOMMENDED PRACTICES

- Data practitioners make every reasonable effort to honor requests from data subjects to be deleted from the dataset (and any and all (re-)distributed copies of the dataset).
- If data subjects revoke their consent and are deleted from the dataset, data practitioners revalidate it to ensure sample representativeness.

GUIDING QUESTIONS

- Have we adhered to data subjects' consent about the storage or deletion of their data at any time?
- How will we honor requests from data subjects to be deleted from the dataset?
 - *When we delete data subjects from the dataset, will we revalidate the dataset for sample representativeness? Are we prepared to return to the Collection phase if the sample is no longer representative?*
- What privacy protocols have we implemented for the stored dataset to prevent data leakages, deanonymization, and other privacy harms?
- Have data subjects and impacted communities given feedback on the storage/deletion protocol?
 - *Does the protocol address community-specific concerns about privacy and erasure?*
- How can we ensure the destruction of all dataset copies across all relevant parties?
 - *Do our dataset distribution policies account for the need to destroy dataset derivatives in the future?*

8 Documentation & Data Governance



Data must be safely stored and managed throughout each phase of a fairness assessment's data lifecycle. The appropriate storage type may differ depending on the dataset, given its particular security vulnerabilities. All decisions at each lifecycle phase must also be documented to ensure replicability and transparency.

KEY RISKS

- Data subject and community-level privacy is compromised due to ineffective data governance.
- Due to inadequate documentation, external and internal stakeholders cannot audit or replicate the data collection and fairness assessment process.
- Documentation is inaccessible or incomplete.

BASELINE REQUIREMENTS

- Data practitioners use strong privacy-preserving techniques throughout the data lifecycle to prevent data leakages and other privacy risks, while not sacrificing data utility to the detriment of statistical minorities in the dataset.
- Data practitioners ensure that collected data and demographics are used in compliance with the informed consent obtained from data subjects throughout the data lifecycle.
- Data practitioners thoroughly document their data collection and fairness assessment process and accessibly archive code and other reproducibility artifacts when possible.

RECOMMENDED PRACTICES

- Data practitioners provide comprehensive documentation of the data collection and fairness assessment process aligned with Partnership on AI [documentation guidelines](#).
- Data practitioners implement feedback and relevant learnings from data subjects throughout the data lifecycle to improve future demographic data collection and fairness assessment processes.

GUIDING QUESTIONS

- What privacy protocols have we implemented to prevent data leakages, deanonymization, and other privacy harms across the data lifecycle?
- Will these protocols adequately protect data subjects and communities, particularly those most at risk of privacy-related harms?
- What are potential privacy weak points in the data lifecycle and how can we mitigate these risks?
- How are we documenting our decisions throughout the data lifecycle?
- Who has access to this documentation?
- Does the governance structure of our data lifecycle allow for the implementation of feedback and relevant learnings at various points?

Acknowledgments

This resource would not be possible without the generous contributions of our Working Group, whose members include: Dr. Amy Dickens (Policy Advisor, Responsible Technology Adoption Unit), Irina Raicu (Director, Internet Ethics Program, Markkula Center for Applied Ethics), Emnet Almedom (Frontline Solutions, formerly with Othering and Belonging Institute, UC Berkeley), Irina Raicu (Director, Internet Ethics Program, Markkula Center for Applied Ethics), Julie Cestaro (New York University), Miranda Bogen, (Center for Democracy & Technology), Nicholas Apostoloff, Nivedha Sivakumar (Apple), Orestis Papakyriakopoulos (Assistant Professor of Societal Computing, Technical University of Munich), Seliem El-Sayed (Ethics Foresight, Google DeepMind), Shamira Ahmed (Executive Director, Data Economy Policy Hub).

We'd also like to thank the following advisors for their invaluable input during the review process: Arjun Subramonian (PhD Student; University of California, Los Angeles), Ariana Aboulafia (Center for Democracy & Technology), Dr. Shyam Krishna (Researcher, The Alan Turing Institute, U.K), Vinay Narayan (Senior Manager, Aapti Institute), Vinhcent Le (Senior Legal Counsel, The Greenlining Institute).

Note that the current version of the Participatory and Inclusive Demographic Data Guidelines is the result of a participatory, iterative multistakeholder process and should not be read as representing views from individual contributors or organizations.

We would also like to thank our PAI colleagues, including Stephanie Bell, Tina Park, Neil Uhl, Rebecca Finlay, Aimee Bataclan, Thalia Khan, Jason Millar, and Stephanie Tsao who made this publication possible.

Version
for Public
Comment

Appendix 1: Glossary

Algorithmic System

Algorithmic systems refer to systems that “ingest problem-relevant information from the environment and produce an action.”⁵ These systems typically depend on large amounts of data and are used in a range of settings, such as healthcare,⁶ hiring,⁷ and education.⁸ Note that these Guidelines focus on static (rather than dynamic) algorithmic systems.

Algorithmic In/justice

Algorithmic justice refers to the design, production, and application of algorithmic systems that enable equal economic, political, and social rights and opportunities for all people. Algorithmic injustice refers to the design, production, and application of algorithmic systems that reproduce, magnify, or create social inequities and harms.

Data Equity

Data Equity is a set of principles and practices to guide anyone who works with data (especially data related to people) through every step of a data project with justice, equity, and inclusivity in mind. Equity is not just an end goal but also a framing for all data work from project funding, motivation, and design to data collection and sourcing, analysis, interpretation, and distribution.

Data Governance

Data Governance refers to the power relations between all the actors affected by how data is collected, accessed, controlled, shared, and used in any given context.⁹ It is also defined as the process of managing the availability, usability, integrity, and security of the data in a system.¹⁰

Data Justice

A concept that refers to justice as how people are made visible, represented, and treated as a result of their production of digital data.¹¹ The fundamental premises of data justice are that data should make visible community-driven needs, challenges, and strengths, represent the community, and treat data in ways that promote community self-determination. Data Equity is a connected but separate concept that refers to “the consideration, through an equity lens, of how data is collected, analyzed, interpreted, and distributed.”¹²

-
- 5 Kochenderfer, Mykel J., Tim A. Wheeler, and Kyle H. Wray. *Algorithms for Decision Making*. MIT Press, 2022.
 - 6 Obermeyer, Ziad, Brian Powers, Christine Vogeli, and Sendhil Mullainathan. “Dissecting Racial Bias in an Algorithm Used to Manage the Health of Populations.” *Science* 366, no. 6464 (October 25, 2019): 447–53. <https://doi.org/10.1126/science.aax2342>.
 - 7 Chen, Le, Ruijun Ma, Anikó Hannák, and Christo Wilson. “Investigating the Impact of Gender on Rank in Resume Search Engines.” In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems - CHI '18*, 1–14. Montreal QC, Canada: ACM Press, 2018. <https://doi.org/10.1145/3173574.3174225>.
 - 8 Rimfeld, Kaili, and Margherita Malanchini. “The A-Level and GCSE Scandal Shows Teachers Should Be Trusted over Exams Results.” *inews.co.uk*, August 21, 2020. <https://inews.co.uk/opinion/a-level-gcse-results-trust-teachers-exams-592499>.
 - 9 Abraham, Rene, Johannes Schneider, and Jan vom Brocke. “Data Governance: A Conceptual Framework, Structured Review, and Research Agenda.” *International Journal of Information Management* 49 (December 1, 2019): 424–38. <https://doi.org/10.1016/j.ijinfomgt.2019.07.008>.
 - 10 “What Is Data Governance and Why Does It Matter?” Accessed February 21, 2024. <https://www.techtarget.com/searchdatamanagement/definition/data-governance>.
 - 11 Dencik, Lina, and Javier Sanchez-Monedero. “Data Justice.” *Internet Policy Review* 11, no. 1 (January 14, 2022). <https://doi.org/10.14763/2022.1.1615>.
 - 12 LATech4Good. “What Is Data Equity and Why Does It Matter?” Accessed January 22, 2024. <https://latech4good.org/news/whatisdataequity>.

Data Leakage

The unauthorized and/or unintended data transmission from within an organization to an external or unauthorized internal party.

Data Misrepresentation

Refers to instances when the demographic categories applied in a dataset do not adequately or accurately represent the identity of the individual being counted.

Data Minimization

Data minimization is a principle that requires data practitioners to limit data collection to what is directly relevant and necessary to accomplish a specific purpose and retain the data only for as long as is necessary to fulfill that purpose.¹³

Data Miscategorization

Refers to instances when an individual is incorrectly classified in a dataset despite the existence of an accurate (representative) data category.

Data Practitioners

Individuals who lead or participate in data collection and/or analysis, such as a fairness assessment process requiring demographic data.¹⁴

Data Subject(s)

An individual, group, or community participating in a data collection process, such as a fairness assessment process requiring demographic data.

Data Reidentification or Deanonimization

Refers to situations where the identity of a person or organization is discoverable even though the individual or organization's name is not available or purposely removed.

Demographic Data

Throughout the literature on algorithmic fairness, demographic data refers to variables used to represent social categories such as gender, race, ethnicity, and sexuality. In our work, we draw from long histories of scholarship that interrogate simplistic categorization schemas and the social harms that can stem from their uncritical adoption. Stemming from these understandings, we see demographic data as an attempt to collapse complex social concepts into categorical variables based on observable or self-identifiable characteristics. While these characteristics are multidimensional, fluid, and often not observable, algorithmic fairness interventions often treat them as singular, self-evident, and stable.

Demographic Inference

Refers to a set of techniques data analysts used to identify and fill in missing demographic traits as accurately as possible by analyzing other available data.

Disparate Impact

Refers to the often-used legal interpretation of “fairness,” which emphasizes determining whether one group experiences different outcomes or treatment (unfair), including when differences emerge unintentionally.

Dynamic Consent

An approach towards consent that allows data subjects to make ongoing, granular decisions about their participation throughout the project.

13 “D | European Data Protection Supervisor,” March 30, 2023. https://edps.europa.eu/data-protection/data-protection/glossary/d_en.

14 data.org. “Accelerate Aspirations: Moving Together to Achieve Systems Change.” Data.Org (blog), January 17, 2023. <https://data.org/reports/accelerate/>.

Equalized Odds

Refers to a commonly used definition of fairness in machine learning (ML) where a model is considered to be operating fairly if data subjects across demographic groups have equal true positive and false positive rates. This means that data subjects from different demographic groups are equally likely to get a positive or negative outcome from the model.

Intersectionality

Intersectionality refers to how systems of social inequality based on gender, race, ethnicity, sexual orientation, gender identity, and other markers of identity overlap and compound to create unique forms of disadvantage. In the context of AI and fairness, [intersectional analysis](#) refers to the consideration of how intersecting forms of social oppression manifest algorithmically, creating harm for certain communities.

Marginalized Groups

Identity groups that face specific (sometimes overlapping and compounding) forms of social, economic, and/or political exclusion and disenfranchisement.

Sociotechnical

An approach in which social structures and technical systems co-inform one another. Assessing just the technical components of a system obscures the human components embedded within them, thereby misrepresenting the consequences and impacts of the system.

Synthetic Data

A dataset algorithmically generated to mimic real-world information, such as demographic or behavioral data. These datasets are typically used for testing and training models.¹⁵

Statistical Minority

Refers to a group within a society that is smaller in size (fewer number of people) than another group. In the case of demographic groups, this may overlap with social minorities, which are defined as groups that experience systematic discrimination, prejudice, and harm based on a demographic trait.

Statistical or Demographic Parity

Refers to a commonly used definition of fairness in machine learning related to the legal doctrine of “disparate impact,” where a model is considered to be operating fairly if each group is expected to have the same probability of experiencing the positive, favorable outcome.

Surveillance Infrastructure

Refers to data-driven tools embedded in the built and/or digital environment that allow for the monitoring of individual behaviors and actions.

Predictive Parity

Refers to a commonly used definition of fairness in machine learning where a model is considered to be operating fairly if the precision rates are equivalent across demographic groups under consideration.

15 Patki, Neha, Roy Wedge, and Kalyan Veeramachaneni. “The Synthetic Data Vault.” In 2016 IEEE International Conference on Data Science and Advanced Analytics (DSAA), 399–410, 2016. <https://doi.org/10.1109/DSAA.2016.49>.