PARTNERSHIP ON AI

WHITE PAPER

Eyes Off My Data

Exploring Differentially Private Federated Statistics to Support Algorithmic Bias Assessments Across Demographic Groups

Sarah Villeneuve
Tina M. Park
Eliza McCullough

Contents

Executive Summary

Designing and deploying algorithmic systems that work as expected every time for all people and situations remains a challenge and a priority. Rigorous pre- and post-deployment fairness assessments are necessary to surface any potential bias in algorithmic systems. As they often involve collecting new user data, including sensitive demographic data, post-deployment fairness assessments to observe whether the algorithm is operating in ways that disadvantage any specific group of people can pose additional challenges to organizations. The collection and use of demographic data is difficult for organizations because it is entwined with highly contested social, regulatory, privacy, and economic considerations. Over the past several years, Partnership on AI (PAI) has investigated key risks and harms individuals and communities face when companies collect and use demographic data. In addition to well-known data privacy and security risks, such harms can stem from having one's social identity being miscategorized or data being used beyond data subjects' expectations. These risks and harms are particularly acute for socially marginalized groups, such as people of color, women, and LGBTQIA+ people. PAI's demographic data work has explored concerns related to data privacy, data security, misuse or abuse of data (including potential discriminatory uses), as well as a more general concern about how quantitative demographic data contributes to (mis) understandings of social identities.

Given these risks and concerns, organizations developing digital technology are invested in the responsible collection and use of demographic data to identify and address algorithmic bias. For example, in an effort to deploy algorithmically driven features responsibly, Apple introduced IDs in Apple Wallet with mechanisms in place to help Apple and their partner issuing state authorities (e.g., departments of motor vehicles) identify any potential biases users may experience when adding their IDs to their iPhones.[1]

In addition to pre-deployment algorithmic fairness testing, Apple followed a post-deployment assessment strategy as well. As part of IDs in Wallet, Apple applied differentially private federated statistics as a way to protect users' data, including their demographic data. The main benefit of using differentially private federated statistics is the preservation of data privacy by combining the features of differential privacy (e.g., adding statistical noise to data to prevent re-identification) and federated statistics (e.g., analyzing user data on individual devices, rather than on a central server, to avoid the creation and transfer of datasets that can be hacked or otherwise misused). What is less clear is whether differentially private federated statistics can attend to some of the other risks and harms associated with the collection and analysis of demographic data. To understand this, a sociotechnical lens is necessary to understand the potential social impact of the application of a technical approach.

This report is the result of two expert convenings independently organized and hosted by

1 IDs in Wallet, in partnership with state identification-issuing authorities (e.g., departments of motor vehicles), were only available in select US states at the time of the writing of this report.

PAI. As a partner organization of PAI, Apple shared details about the use of differentially private federated statistics as part of their post-deployment algorithmic bias assessment for the release of this new feature.

During the convenings, responsible AI, algorithmic fairness, and social inequality experts discussed how algorithmic fairness assessments can be strengthened, challenged, or otherwise unaffected by the use of differentially private federated statistics. While the IDs in Wallet use case is limited to the US context, the participants expanded the scope of their discussion to consider differential private federated statistics in different contexts. Recognizing that data privacy and security are not the only concerns people have regarding the collection and use of their demographic data, participants were directed to consider whether differentially private federated statistics could also be leveraged to attend to some of the other social risks that can arise, particularly for marginalized demographic groups.

The multi-disciplinary participant group repeatedly emphasized the importance of having both pre- and post-deployment algorithmic fairness assessments throughout the development and deployment of an AI-driven system or product/feature. Post-deployment assessments are especially important as they enable organizations to monitor algorithmic systems once deployed in real-life social, political, and economic contexts. They also recognized the importance of thoughtfully collecting key demographic data in order to help identify group-level algorithmic harms.

The expert participants, however, clearly stated that a secure and privacy-preserving way of collecting and analyzing sensitive user data is, on its own, insufficient to deal with the risks and harms of algorithmic bias. In fact, they expressed that such a technique is not entirely sufficient for dealing with the risks and harms of collecting demographic data. Instead, the convening participants identified key choice points facing AI-developing organizations to ensure the use of differentially private federated statistics contributes to overall alignment with responsible AI principles and ethical demographic data collection and use.

This report provides an overview of differentially private federated statistics and the different choice points facing AI-developing organizations in applying differentially private federated statistics in their overall algorithmic fairness assessment strategies. Recommendations for best practices are organized into two parts:

1. General considerations that any AI-developing organization should factor into their post-deployment algorithmic fairness assessment
2. Design choices specifically related to the use of differentially private federated statistics within a post-deployment algorithmic fairness strategy

The choice points identified by the expert participants emphasize the importance of carefully applying differentially private federated statistics in the context of algorithmic bias assessment. For example, several features of the technique can be determined in such a way that reduces the efficacy of the privacy-preserving and security-enhancing

aspects of differentially private federated statistics. Apple's approach to using differentially private federated statistics aligned with some of the practices suggested during the expert convenings: the decision to limit the data retention period (90 days), allowing users to actively opt-into data sharing (rather than creating an opt-out model), clearly and simply sharing what data the user will be providing for the assessment, and maintaining organizational oversight of the query process and parameters.

The second set of recommendations surfaced by the expert participants primarily focus on the resources (e.g., financial, time allocation, and staffing) necessary to achieve a level of alignment and clarity on the nature of "fairness" and "equity" AI-developing organizations are seeking for their AI-driven tools and products/features. While these considerations may seem tangential, expert participants emphasized the importance of establishing a robust foundation on which differentially private federated statistics could be effectively utilized. Differentially private federated statistics, in and of itself, does not mitigate all the potential risks and harms related to collecting and analyzing sensitive demographic data. It can, however, strengthen overall algorithmic fairness assessment strategies by supporting better data privacy and security throughout the assessment process.

Introduction

Many organizations are committed to developing and releasing AI-driven products and features that are both inclusive of a broad base of users and function effectively across a diversity of users. To ensure this, fairness assessments are used to identify and mitigate potential algorithmic bias, especially bias that might be experienced by historically and currently marginalized demographic groups, such as gender non-conforming and transgender individuals, people of color, people with disabilities, and women. Bias in algorithmic systems[2] can occur for a number of reasons and is often the result of the system making correlations or establishing trends that have the effect of discriminating across groups, even if that is not the intention or purpose.

2 See Appendix 4 for a more detailed definition.

AI-developing organizations can take steps to test algorithmically driven systems, products, and features for bias before they are released (pre-deployment). However, pre-deployment testing cannot identify all possible issues so it is important to conduct post-deployment analysis to determine if users are experiencing any biased, or otherwise negative, interactions or outcomes. Most current algorithmic fairness techniques, whether pre-deployment or post-deployment, require access to sensitive demographic data[3] (such as age, ethnicity, gender, and race) to make performance comparisons and standardizations across groups. Post-deployment algorithmic fairness assessments frequently rely on the collection of new user data, as opposed to the use of existing datasets, as it is an opportunity to observe the algorithmic system in use by real people. However, AI practitioners face several challenges trying to procure the data necessary to identify and understand the nature of the bias in their algorithmic systems.

3 The term "demographic data" refers to information that attempts to collapse complex social concepts into categorical variables based on observable or self-identifiable characteristics, such as gender, race, or ethnicity.

The Challenges of Algorithmic Fairness Assessments

Organizations that develop algorithmic systems are faced with competing consequences. On the one hand, organizations are eager to ensure their products and systems perform fairly and as expected for all users. Collecting and analyzing user data, alongside key demographic characteristics, is necessary to ascertain whether certain groups of people are not receiving a fair and high-fidelity experience, indicating potential algorithmic bias. On the other hand, due to a long history of discriminatory behavior enabled by the collection and use of demographic data, organizations face regulations and other restrictions.

As previous Partnership on AI (PAI) research has highlighted, the collection and use of demographic data is entwined with highly contested social, political, and economic considerations. Individual privacy and anti-discrimination laws (e.g., the Civil Rights Act of 1964 and the Fair Housing Act in the United States and the General Data Protection Regulation in the European Union) have restrictions related to the collection of demographic data. In some instances, organizations may be disincentivized from

exploring potential algorithmic bias, as they may face legal consequences if they know of discriminatory or biased behaviors without having a plan to address and mitigate them. However, simply choosing to not collect pertinent demographic data to avoid such responsibility — often referred to as "fairness through unawareness" — obscures the discriminatory impacts of algorithmic systems and can contribute to the perpetuation of social inequities faced by marginalized communities.

An individual's demographic characteristics could also be used to re-identify a specific individual, revealing their other user data, including behavioral data, resulting in the loss of privacy.[4] Additionally, the collection and use of sensitive and fine-grained individual user data for advertising has resulted in racially targeted misinformation campaigns, predatory lending, and the loss of public trust. Possible re-identification may result in individuals being targeted or otherwise surveilled based on specific demographic characteristics, further expanding and enabling surveillance infrastructures.[5] Inadequately designed data categories and models have contributed to empirical narratives that reify and deepen social stereotypes, which in turn cause harm to socially marginalized communities. Yet socially marginalized communities have also argued for the collection of demographic data, as such data is integral for identifying discriminatory behaviors and outcomes.[6]

4 The threat of reidentification of specific individuals is particularly relevant for cases in which corporate data is requested by and made available to state agencies. See: Leetaru, K. (2018, July 20).

5 See Appendix 4 for a more detailed definition.

6 The following examples describe how and why marginalized communities collect and leverage demographic data to advance equity: use of data for labor organizing (Bottom-Up Organizing with Tools from On High: Understanding the Data Practices of Labor Organizers), alternative data collection and use methods by Indigenous communities (Indigenous Data Sovereignty), localized data collection efforts to inform city-level policies and practices (Our Data Bodies: Reclaiming Our Data).

Prioritization of Data Privacy: An Incomplete Approach for Demographic Data Collection?

Several privacy-preserving techniques have been proposed that seek to address some of the concerns posed by demographic data collection and analysis.[7] In general, these privacy-preserving techniques work to ensure individual privacy through anonymization and limiting how an individual's information can be accessed and analyzed to reduce re-identification risks.

One such technique is differentially private federated statistics, which combines two approaches: differential privacy and federated statistics (also referred to as federated learning[8]). Differential privacy refers to an approach in which random statistical noise is added to data to enforce privacy constraints. Federated statistics involves running local computations on an individual's device and only making the composite results (rather than the specific data from a particular device) visible at a central or external level.[9]

It has been shown that these two techniques can be designed and implemented together[10] to ensure the privacy of individuals' sensitive data. Sensitive user data can be collected and analyzed on an individual's device to determine whether the individual is experiencing algorithmic bias (federated statistics). Statistical noise can be introduced to the output data and ultimately shared with the organization assessing the algorithm (differential privacy) to ensure sensitive user information is protected against re-identification. While the application of differentially private federated statistics in the context of algorithmic

7 See Appendix 3 for a more detailed definition.

9 See the section below titled "Differentially Private Federated Statistics" for a more detailed explanation of both differential privacy and federated statistics.

10 See also: Federated Learning with Formal Differential Privacy Guarantees

fairness is relatively new, it is viewed as a promising post-deployment technique for overcoming some of the barriers and challenges related to the collection and use of sensitive user data.

However, it is important to acknowledge that data privacy and security are not the only factors that concern individuals whose data is being used and responsible AI advocates. For example, there are broader questions about whether the appropriate features of social identity and interactions with algorithmic systems are being measured and studied for the purposes of algorithmic fairness. It cannot be assumed that privacy-preserving statistical approaches are inherently designed to also grapple with these other fairness questions. In this report, we explore how differentially private federated statistics can be best leveraged to analyze algorithmic systems for potential bias, examining the potential limitations and negative implications of its use.

Premise of the Project

In order to understand the sociotechnical dimensions of various approaches to algorithmic bias and fairness assessments, Partnership on AI (PAI) relies on its multistakeholder model to identify both technical considerations and potential social risks and impacts. PAI convenes experts from different sectors and disciplines, ranging from technologists (technical experts) to social scientists (social issue experts) to civil society advocates (social impact experts) to provide a more holistic understanding of a given algorithmic issue. For example, PAI considers algorithmic fairness not only as the pursuit of statistical parity,[11] but as the pursuit of social equity that is attentive to structural inequities, power asymmetries, and histories of discrimination and oppression.

11 See Appendix 4 for a more detailed definition.

PAI also recognizes that responsible AI principles are challenging to operationalize. Strategies to contend with algorithmic harms may be well-considered on paper, but run into obstacles when being implemented within an organization. There may be legal and organizational considerations, from issues of legal liability to the necessary staffing and organizational incentives, that may stymie the implementation of responsible AI practices within an organization. Furthermore, an algorithmic system is not deemed to be responsibly or ethically developed simply because it functions as intended. It is also necessary to take into account the way it was developed (e.g., development process) and the components that went into its creation (e.g., datasets). As they say, "the devil is in the details" and these more quotidian development and design decisions are often the choice points not discussed in responsible AI guidance. For these reasons, the opportunity to observe and learn directly from teams and organizations as they implement various responsible AI practices or strategies is invaluable.[12]

12 See Appendix 1 for more information about Partnership on AI's Fairness, Transparency, and Accountability (FTA) program area and its existing work on the use of demographic data for algorithmic fairness purposes.

To better examine the potential of differentially private federated statistics to support a

more robust approach to algorithmic bias identification, PAI collaborated with a major technology company from its Partner community.[13] As part of their roll-out of IDs in Wallet in the United States, Apple implemented differentially private federated statistics to support their post-deployment algorithmic fairness assessment strategy.[14]

PAI organized two multistakeholder expert convenings, using the details by Apple to host a more grounded and specific discussion about differentially private federated statistics in a real case.[15] Each three-hour virtual workshop was organized to examine how this data privacy mechanism can support or limit more responsible data collection and analysis for algorithmic fairness assessments.[16] The context of US digital identification cards — particularly the stakes of ensuring that all people are able to successfully onboard and utilize a digital identification card if they wish to — surfaced key points about the importance of addressing algorithmic bias using tools like differentially private federated statistics. This included discussions about how different social identities are defined and measured in order to determine whether any experience of group-level algorithmic harm related to that social identity; or whether people with highly marginalized social identities would feel safe disclosing their identities, even for the purposes of identifying potential algorithmic harm.

The 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.[17] These convenings were designed to explore differentially private federated statistics through both social and technical lenses. Participants were also encouraged to consider the risks surrounding sensitive demographic data collection and analysis that may not be fully mitigated through the application of differentially private federated statistics and the additional steps organizations could take to strengthen their overall algorithmic fairness approach.

This white paper is based on the insights provided by the multistakeholder body of experts across the two convenings, as well as a review of available secondary literature on differential privacy, federated learning, and the social considerations of demographic data collection and use. Regular, weekly discussions with key Apple staff involved with the implementation of differentially private federated statistics for algorithmic bias identification in IDs in Apple Wallet[18] also helped to clarify PAI's understanding of Apple's overall approach to algorithmic fairness.

13 See the section titled "Funding Disclosure" for more information regarding Partnership on AI's relationship with Apple, Inc.

14 See Appendix 2 for more details about Apple's algorithmic fairness assessment strategy for their new IDs in Wallet feature.

15 The purpose of the workshop and this report is to provide the AI community with guidance on an important and novel technique. While Apple benefits from the case-specific discussion hosted by Partnership on AI, the role of PAI — and the experts who participated in the convenings — is not to assess Apple on the relative success (or lack thereof) of a technique they chose to employ in the roll-out of their IDs in Wallet feature.

16 See Appendix 3 for more details about the PAI-sponsored expert convenings.

17 Although the case study provided by Apple is specific to the United States, also included in the multistakeholder convenings were experts from Canada and the United Kingdom who noted considerations for how use of differentially private federated statistics in their socio-political contexts may be similar or different. Additional research should be conducted to determine how use of differentially private federated statistics may differ in non-Western contexts and by non-corporate organizations developing and/or deploying AI.

18 This included regular touchpoints with the Product and Engineering teams, as well as some consultations with the Business Development, Marketing, and Legal teams.

A Sociotechnical Framework for Assessing Demographic Data Collection

This paper provides a *sociotechnical*[19] examination of differentially private federated statistics, bringing these concerns of individual and community-level social risks and harms alongside considerations of technical accuracy. To do so, this paper looks at the use of differentially private federated statistics within the context of a broader algorithmic fairness assessment strategy undertaken by a team or organization, rather than assessing differentially private federated statistics independent of this context. This type of analysis reveals that when integrating differentially private federated statistics, there are a number of design choices organizations and teams must make to ensure the overall strategy to collect and use sensitive demographic data is conducted responsibly. If organizations neglect to think critically about the various design choices provided to them when using this technique, they risk further entrenching historical discrimination and introducing new forms of bias, rendering their bias mitigation efforts moot. Based on these findings, we provide recommendations for organizations interested in using this technique to achieve their fairness goals.

19 See Appendix 5 for a more detailed definition.

The challenges individual AI developers and organizations face when attempting to conduct an algorithmic fairness assessment have been enumerated by PAI and other AI ethics researchers. Privacy-preserving techniques, such as differentially private federated statistics, hold the promise of addressing some of these challenges so AI developers may collect the necessary demographic data to identify potential algorithmic bias. For example, by obfuscating the link between a user and their demographic data, reducing the likelihood of re-identification, and minimizing opportunities for data breaches, differentially private federated statistics allow data collection and analysis to be aligned with legal and regulatory requirements protecting individual data privacy. Such protections also lessen the risks to organizations by reducing the likelihood of data mismanagement which often results in reputational damage, erosion of consumer trust, and significant legal implications.

However, other concerns and risks remain. For example, at an organizational level, data privacy preservation does not inherently mitigate challenges with identifying the most appropriate demographic data categories to model and capture the potential algorithmic bias (selecting appropriate demographic measurements). The psycho-social risks of datafication,[20] both at the individual and the community level, are also not fully and automatically resolved through the application of privacy-preserving data techniques. Misrepresentation[21] and miscategorization[22] can result in psychological harm for individuals, causing them to feel as if they need to alter their behavior or appearance

20 21 22 See Appendix 4 for more detailed definitions.

in order to "fit the mold" of the category with which they identify. The likelihood of re-identification may be significantly reduced, thereby minimizing the risk of being targeted for specific social identity markers. However, those contributing their data (data subjects), particularly those of marginalized social identities, may be asked to contribute even more data tracking even greater minutia of their lives in the name of algorithmic fairness assessment. This expanded data collection may unintentionally result in increased surveillance if the data collected is not protected from misuse or undisclosed uses.

Any proposed innovation to the collection and use of demographic data should be assessed through a sociotechnical framework because these kinds of risks and harms are not always evident when assessed through a single risk or harm factor. Risks related to consumer database breaches may limit what an organization chooses to collect, store, and analyze, even if for algorithmic fairness purposes. However, avoidance of any demographic characteristics may lead to the inability to assess for bias (e.g., "race-blind" algorithms). For this reason, a robust algorithmic fairness strategy that involves the collection and use of demographic data is defined as one that overcomes the organizational and legal barriers while also mitigating social risks. A summary of these challenges and risks associated with demographic data collection and analysis is provided in Table 1 by which we consider differentially private federated statistics.

TABLE 1: Challenges and Risks Associated with Demographic Data Collection and Analysis[23]

Organizational Concerns	Legal Barriers	Social Risks to Individuals	Social Risks to Communities
• Organizational priorities • Public relations risk • Discomfort (or lack of expertise) with identifying appropriate demographic groups	• Anti-discrimination law • Privacy policies	• Unique privacy risks associated with the sharing of sensitive attributes likely to be the target of fairness analysis • Possible harms stemming from miscategorizing and misrepresenting individuals in the data collection process • Use of sensitive data beyond data subjects' expectations	• Expansion of surveillance infrastructure in the name of fairness • Misrepresenting and miscategorizing what it means to be part of a demographic group or to hold a certain identity • Data subjects ceding the ability to define for themselves what constitutes biased or unfair treatment

23 Appendix 5 provides an expanded version of this table with a more detailed definition and illustrative examples. See "'What We Can't Measure, We Can't Understand': Challenges to Demographic Data Procurement in the Pursuit of Fairness" and "Fairer Algorithmic Decision-Making and Its Consequences: Interrogating the Risks and Benefits of Demographic Data Collection, Use, and Non-Use" for a complete description of challenges and risks associated with demographic data collection and usage.

Differentially Private Federated Statistics

Differentially private federated statistics is a privacy-preserving technique that combines two approaches, differential privacy and federated statistics, in order to enable large-scale, interactive data analysis while helping to ensure that an individual's sensitive information remains private and secure. Both differential privacy and federated statistics can be implemented independent of one another. In this section, we explore these two approaches first individually, and then as complementary techniques, before discussing the motivations for combining them in support of data analysis efforts.

Differential Privacy

As Cynthia Dwok and Aaron Roth write, "[d]ifferential privacy addresses the paradox of learning nothing about an individual while learning useful information about a population." Differential privacy, first defined in 2006, is not an algorithm or fixed system, but rather an analytical approach that can be constructed in various ways with the common aim of preventing a set of (anonymized) personal data from being re-identified. This data processing framework proposes to provide a strong privacy guarantee to individuals by enforcing privacy constraints locally (e.g., on an individual's device) or centrally (e.g., the server after data has been collected from individuals) by adding random statistical noise. This infusion of random statistical noise makes it difficult to identify any given individual who has contributed their data. The noise, however, is designed to not impede group-level analysis. Essentially, differential privacy works to provide data analysts with trends or patterns across groups as opposed to individual-level information.

This approach addresses organizational and legal concerns surrounding privacy held by organizations, particularly related to the consequences of accidentally revealing private user data, while allowing them to obtain insights about their users or consumers. It also eases concerns held by those contributing their data (data subjects), as the risk of re-identification or losing their anonymity may be minimized.

Differential privacy has been most commonly used in instances where organizations, such as government agencies or healthcare providers, wish to publish datasets while maintaining the privacy and confidentiality of those who contributed their data. Many companies, including Google, Meta, and Apple, also employ this technique when collecting data from their users. More recently, differential privacy has received attention due to a debate sparked by its use by the US Census Bureau. This debate focused mainly on the tradeoff between privacy and accuracy, and the impact of this tradeoff for marginalized groups. Critics of differential privacy argued that while differential privacy ensures strong privacy protections against de-anonymization, the infusion of statistical noise can reduce

the accuracy of analyses performed on the dataset particularly for groups that make up a statistical minority.[24]

24 See Appendix 4 for a more detailed definition.

Federated Statistics

Federated statistics (drawn from the more commonly known federated learning) is a machine learning (ML) technique that enables organizations to access and use data from multiple, discrete devices without the need to collect and store this data in a centralized database. In doing so, federated statistics provides some privacy protection and data security, since any personal data used to train an ML model does not have to leave an individual device and be at risk of data breaches — either during data transfer or due to being stored on a central server. This technique is also scalable as it allows for multiple organizations to collaborate on a given ML task without requiring them to share large volumes of raw data with each other. In the case of algorithmic fairness, this allows for bias assessments to take place across a large volume and wide array of users to understand the full impact of any potential bias issues.

Federated statistics necessitates the use of discrete, individual systems or devices where analysis can be performed. Limiting factors for this technique, therefore, include the computational capabilities, storage, network connectivity, and power of these devices. Federated learning, on its own, is also widely known to be vulnerable to privacy and security issues since the data provided to the central server can be used to identify individuals.[25] This technique is also susceptible to model poisoning attacks[26] since malicious users can directly influence the global model by infusing incorrect or messy data from their device.

25 Lyu, Lingjuan, Han Yu, and Qiang Yang. "Threats to Federated Learning: A Survey." arXiv, March 4, 2020. http://arxiv.org/abs/2003.02133.

Federated statistics have been most commonly used to train ML models. For example, AI-developing organizations can send a centralized ML model (trained on publicly available data or untrained entirely) to each individual device, and each device will train a copy of that model locally before sending back the training results to a central server where results are aggregated and the centralized model is updated. In the context of algorithmic fairness problems, instead of sending large volumes of user data to a central server, data scientists can send queries (specific questions that can be answered through data) to individual devices and receive an aggregate report back that allows them to identify trends across groups.[27]

26 See Appendix 4 for a more detailed definition.

27 It should be noted that as of the writing of this white paper, there is no formal definition for federated algorithms and thus other scholars' definitions may slightly vary.

Differentially Private Federated Statistics

Combining differential privacy and federated statistics allows for large-scale, interactive data analysis while ensuring an individual's sensitive information remains private and secure. Data scientists are able to gain insight into aggregate trends by sending queries to each data source (federated statistics). By adding statistical noise to the data, analysts can ensure individual privacy. Additionally, a secure aggregation protocol[28] can also be used to

28 See Appendix 4 for a more detailed definition.

ensure that only the aggregate of individual reports generated from each query is visible to those conducting the data analysis (differential privacy). In other words, this technique ensures that no raw individual data is stored or visible to data analysts at the central level, restricting visibility to only the aggregate of reports generated.

One of the key strengths that differentially private federated statistics has over other privacy-enhancing techniques, such as Secure Multi-Party Computation (SMPC),[29] is its ability to scale across numerous external actors, vendors, and contexts. This is particularly useful for an organization that is seeking to assess the fairness of a system that operates in multiple and highly varied social, political, and economic contexts as highly localized and specific trends can be identified.

29 See Appendix 4 for a more detailed definition.

Like federated statistics, differentially private federated statistics require specific infrastructure in order to operate: discrete devices where an individual's data is collected, stored, and analyzed. Conducting such analyses, depending on their complexity, may require a high local memory capacity and a large amount of computing power. Internet connection is needed for the local device to transfer its data reports to a central server and for queries to be sent to the devices. These hardware requirements may limit the types of AI/ML organizations that are able to take advantage of differentially private federated statistics as part of their algorithmic fairness assessment strategy.

While layering these two approaches effectively mitigates the privacy and security vulnerabilities of federated statistics, the use of federated statistics does not address the challenge of tradeoffs inherent to differential privacy, namely the loss of data utility due to increased privacy. In the following discussion, we review some of the key considerations for the design of a robust algorithmic fairness assessment which relies on the use of differentially private federated statistics as a post-deployment technique to preserve the privacy of individual users while collecting and analyzing demographic data.

A Sociotechnical Examination of Differentially Private Federated Statistics as an Algorithmic Fairness Technique

Differentially private federated statistics is designed to address data privacy concerns. However, as an approach, it was not explicitly developed for the context of algorithmic fairness assessments. As such, it is not inherently designed to address the other social risks associated with collecting and using demographic data, such as miscategorization or reinforcement of oppressive categories.[30] However, this does not mean that differentially private federated statistics pose barriers to the efficacy of other strategies and approaches for responsible and ethical data collection analysis. Rather, when using differentially private federated statistics to enhance the privacy of fairness assessments, there are a number of design choices and variables to consider. It is important to note, however, that such choices may result in trade-offs, for example between privacy and analytic accuracy.[31]

30 For example, these harms could include the group-level miscategorization of gender non-conforming individuals as male or female, reinforcing the oppressive category of the gender binary.

31 See Appendix 4 for a more detailed definition.

As previously mentioned, the ability of this technique to support a successful fairness assessment (e.g., one that overcomes the organizational and legal barriers and mitigates social risks) relies on both the design choices made by the organization and the conditions in which it is implemented. Below we detail a number of sociotechnical considerations that influence the efficacy of a fairness assessment strategy in terms of protecting individual user privacy and identification of potential algorithmic bias, particularly for socially marginalized groups.

While conducting research for this project, expert convening participants frequently discussed the importance of a full pre- and post-deployment algorithmic fairness assessment strategy, especially if the aim is to mitigate the harms experienced by socially marginalized communities. To capture the advantages presented by differentially private federated statistics, organizations must consider the way the problem is defined, the way that data is collected, who continues to be excluded and "not seen" in the data, and how the data is interpreted to guide decision-making on algorithmic bias mitigation to prevent risks and harms for individuals and communities.

General Considerations for Algorithmic Fairness Assessment Strategies

The considerations are divided between two broad categories: 1) those related to the overall design of an algorithmic fairness assessment strategy, and 2) those related to the specific design of a differentially private federated statistics approach within that fairness

assessment strategy. The first set of considerations should be understood as some of the fundamental components needed in a broader algorithmic fairness assessment strategy. Poorly determined decisions among this set of considerations may weaken the efficacy of any attempt to identify algorithmic bias, whether or not differentially private federated statistics is applied. The second set of considerations is specific to how the differentially private federated statistics is applied so that its advantages (e.g., privacy preservation) are maximized.

In general, algorithmic fairness assessments benefit from being treated as an organizational priority, as they require additional expertise, time, and organizational incentives to implement as part of the development of an algorithmic system. This is especially the case for resource-intensive approaches like differentially private federated statistics which may require additional computing capacity and team members capable of administering the overall application of differentially private federated statistics.

Algorithmic systems should be tested for bias at multiple points of development and deployment. Pre-deployment testing and assessments provide assurances that algorithmic systems have been thoroughly vetted to minimize any harmful impacts once they are in operation and interacting with the general public. Post-deployment assessments allow organizations to monitor algorithmic systems as they operate in complex human contexts. Doing so not only protects additional people from being harmed or otherwise negatively impacted by the algorithmic system but allows for developers to improve and innovate on existing algorithmic models. In order to leverage the learnings from post-deployment algorithmic fairness assessments, organizational processes must be in place to incentivize or require teams to address and resolve any identified issues.

Because algorithmic bias is sociotechnical in nature, non-technical experts such as subject matter experts in social inequality and qualitative researchers are important members of algorithmic fairness assessment teams. As we will discuss further, they are especially advantageous when defining relevant demographic groups, choosing data collection methods, considering the balance of privacy and accuracy relevant to different social groups, defining appropriate fairness or bias frameworks to apply to models, and other components of the fairness assessment process. Organizations might also work in collaboration with community groups, particularly those that advocate on behalf of marginalized communities, to support these aspects of algorithmic fairness assessment. Ultimately, effective collaboration with external experts and groups requires clear communication, establishment of trust, potential compensation (monetary or in-kind), and organizational structures to support ongoing engagement. While resource-intensive, these collaborations can lead to more robust, inclusive, and equitable fairness assessments.

RECOMMENDATION(S)

- Team members involved in conducting the overarching fairness assessment, (of which differentially private federated statistics is one component) should include meaningful engagement with non-technical experts and community groups to inform their overall approach.
- This should involve setting expectations, maintaining communication, and providing compensation for those external to the organization who contribute their time and expertise.
- Organizations should provide teams with adequate time and resources to design and deploy algorithmic fairness assessments.
- Teams should obtain executive, leadership, and middle-management buy-in to ensure they receive the proper support to effectively address any bias identified.

Defining Fairness

Before beginning any type of fairness assessment (pre- or post-deployment), an organization should work towards aligning social and statistical definitions of fairness they are employing for that specific assessment. As noted by many social scientists and sociotechnical experts, fairness is an "essentially contested concept," meaning it has "multiple context-dependent, and sometimes even conflicting, theoretical understandings." To further complicate the definition and assessment of fairness and bias, there are many ways to translate understandings of fairness or bias into a set of statistical measurements. For example, "disparate impact" is one interpretation of bias used in legal, and increasingly in algorithmic, contexts. It is defined as a situation which appears neutral (e.g., everyone has the same odds or outcomes), but one group of people (usually of a protected class or characteristic) is adversely affected or implicated. A common measurement of disparate impact is the "80% rule,"[32] which sets a quantified threshold for discriminatory or biased outcomes. Other frequently used statistical approaches to fairness include predictive parity[33] and demographic parity[34] (also known as statistical parity) when assessing systems for disparate impact. These technical definitions attempt to collapse social definitions of unfairness into mathematical formulas in very specific ways and have varying benefits and challenges depending on the context of the application.

Given the many ways fairness (or bias) can be defined and interpreted mathematically, it is important for organizations to both 1) apply interpretations and measurements of fairness that align with their overall values as an organization; and 2) share and discuss how fairness is being defined, both technically and socially, with the public (e.g., other researchers, policymakers, their users). It may not be possible (or worthwhile) to assess for all interpretations of fairness, so maintaining transparency about which notion of fairness is being pursued — and why — can help teams and organizations navigate discussions and criticisms related to their algorithmic fairness assessment strategies.

For example, an organization using the 80% rule or predictive parity should not claim that

they are attempting to ensure fairness for all users, as these approaches do not ensure that all users will be treated equally — or experience similar outcomes — across groups, only that it will perform fairly for a majority of users. It is possible for organizations to apply many different measures of fairness in order to triangulate towards a broader interpretation of fairness. For example, an organization may rely on the 80% rule as a starting point for their fairness assessment, helping them to flag performance areas that require more granular analysis.

Who gets to define fairness for algorithmic fairness assessment strategies is also important to consider. There is increasing interest in engaging outside experts, members of adversely affected communities, broad user bases, and the general public as part of the development process. Algorithmic bias can have severe impacts on an organization's reputation, so engaging those who will be impacted by the end product, including in the defining of "fairness," is one proposed way of grappling with algorithmic bias, expanding public participation while building better products and systems enjoyed by users. Again, given the complexity of defining and measuring fairness, it may be appropriate to adopt a "mixed methods" approach, using both survey and qualitative methods like community-based interviews or focus groups. Being able to respond to these expectations with clear communication about what type of "fairness" the organization is trying to achieve through a fairness assessment will go far in building trust and mitigating miscommunication and harm.

RECOMMENDATION(S)

- Organizations should seek alignment between technical (e.g., statistical) and non-technical (e.g., sociological) definitions of fairness.[35]
- Organizations should seek alignment between developer and user or public understanding and measurement of fairness.
- Organizations should practice transparency when it comes to how they define fairness.

Defining Relevant Demographic Categories

Just as fairness is an "essentially contested concept," demographic categories are contested and context-specific. Which characteristics are measured (e.g., gender versus sex or race versus ethnicity) and the categories that are provided (e.g., male/female, cisgender woman/transwoman/gender non-binary, or female/male/two-spirit) change across different assessment situations based on such factors as what the algorithmic system does and where it is deployed.

For any algorithmic bias assessment, it is critical that the salient demographic categories are identified and appropriately defined for the analysis. This comes down to understanding which demographic categories are salient axes of bias for the context in which their product or system is operating in. In terms of defining demographic categories, some

35 Technical approaches to fairness mean statistically defined fairness, such as the statistical probability of outcome A being the same as outcome B (a 50-50 "chance"). Non-technical or sociological approaches to fairness could include something such as a sliding scale for payment, based on income, such that those who are financially poorer pay less than their financially richer counterparts (recognizing that spending $10 if you earn $20 a day can have a substantially greater negative impact on your overall financial well-being than for someone who earns $1000 a day).

organizations may look to existing government taxonomies (for example, in a US context, the racial and ethnic categories provided in the US census may be viewed as a starting point to select salient social identity categories). However, given the known limitations of government-defined taxonomies, it is also important to conduct research to identify alternative categories and/or redefine categories to map against how they are understood and used in specific socio-political contexts.

Much like tapping into expertise outside the development team to define "fairness," socio-political experts, users, and members of the general public can help refine the selection of appropriate demographic categories to assess algorithmic fairness. The categories of data used to measure or assess possible bias is a particularly sensitive (and contested) area of concern for many members of marginalized social groups, as the way individuals are measured may generate other harm to these marginalized groups. These harms can include miscategorization (when an individual is misclassified despite there being a representative category that they could have been classified under) or misrepresentation (when categories used do not adequately represent the individual as they self-identify). Not only can miscategorization or misrepresentation lead to psychological and emotional harm via feelings of invalidation and rejection, but entire groups of individuals would be rendered invisible within the data because they are effectively not being counted. This is particularly of concern, as important decisions are often made based on statistical analyses of populations, such as political representation, allocation of social services, and functionality of products and features.

RECOMMENDATION(S)

- Organizations should allocate the necessary resources to conduct original research on appropriate measurements and metrics for the fairness assessment process.
- In order to yield a more inclusive fairness assessment process, this may include exploring what demographic categories are relevant to the deployment of their systems or features, as well as how different communities may define their own demographic attributes.

Data Collection

There are important considerations related to how an individual is able to offer the information necessary for the algorithmic fairness assessment. There are a number of methods for data collection organizations could employ, each with their own advantages and disadvantages. For example, an individual can be directly asked to select their gender identity from a limited list to determine if an algorithmic system is behaving in a gender-biased manner. These may provide the algorithmic assessment team with a "clean" dataset ready for analysis, but individuals — especially those who feel they are not accurately being reflected in the categories offered — may choose not to volunteer their information for assessment. Generally, individuals from social groups that experience miscategorization

or violence due to their minority identity may be less likely to volunteer demographic and other private information out of concern for their safety. For example, in the United States where immigrant documentation status is highly politicized and policed, researchers found that asking about citizenship status on the US census significantly increased the percent of questions skipped and made respondents less likely to report that members of their household were Latino/Hispanic. For this reason, datasets would be incomplete across individuals who are very likely to be underrepresented in other datasets (e.g., training, testing) and experiencing regular exclusion from technological advances.

Another popular method is the use of open forms which allow for individuals to self-report or self-identify without the requirement to select from predefined categories. This technique can result in incomplete or highly variable datasets requiring resource-intensive cleaning which can impact the accuracy of the fairness assessment. However, it can also provide individuals with autonomy over how they define themselves which is particularly crucial for communities whose identities have been historically excluded from formal data collection efforts.

Organizations may also rely on demographic inference methods[36] to proxy for traits, such as using computer vision technology to algorithmically ascribe skin tone as a proxy for race or ethnicity. While this technique can provide organizations with robust, uniform datasets, it removes agency from individuals in their ability to self-identify and can lead to miscategorization or misrepresentation. Additionally, existing measurements have been critiqued, and in some cases abandoned, for the limitations they pose for marginalized groups. For example the Fitzpatrick scale, which is often used to classify skin tone, has received criticism for its failure to accurately capture darker skin tones. In 2022, Google implemented the Monk scale which expands on the skin tone shades first established by the Fitzpatrick scale to improve categorization accuracy, particularly for people with darker skin tones. However, even this more inclusive scale fails to prevent all instances of miscategorization or overcome the issue of self-identification.

36 See Appendix 4 for a more detailed definition.

A separate issue is one of individual user consent for the collection and use of their private data. There have been a number of well-documented cases where individual data has been taken and used for other uses, such as targeted political advertising, without the individual's knowledge or consent. To maintain public trust, as well as stay in accordance with different data privacy laws, it is important that when collecting personal data, active and knowledgeable consent is received by the individual providing the data. In order to receive knowledgeable consent,[37] it may be necessary to acquire ongoing consent, rather than consent at a single point in time to provide broad accessibility to an individual's data. Organizations should provide individuals with the opportunity to accept or refuse the provision of their data, alongside information on how their data might be used. Accessibility, clarity, scope, frequency, and language are all important elements to consider when designing an ethical consent process. For example, for IDs in Apple Wallet, users are asked

37 See Appendix 4 for a more detailed definition.

to offer the use of their data for algorithmic bias assessment at the time of onboarding but may opt out at any time. If users do opt in, data is only accessed and used for 90 days (from the time of opt-in). Convening participants noted that having simple mechanisms to opt out and having a clear data retention period are important features to maintain.

Differentially private federated statistics present an advantage for some of these data collection concerns, as it ensures that the collection, storage, analysis, and sharing of an individual's demographic data is more likely to remain anonymous and less likely to be unknowingly taken, or reconstructed, by another party. By keeping an individual's information on their own device, rather than storing it on a central database that might be less secure, concerns regarding data breaches and misuse of user databases may be allayed. By introducing noise to the data and only reporting out results (as opposed to raw data), individuals are more protected from having their identities de-anonymized or reconstructed, mitigating some concerns related to surveillance or targeted violence.

The additional protections offered by differentially private federated statistics can also make it possible to deviate from government-defined demographic categories, as individuals may respond in ways that are more reflective of their own perceived identities without fear of being marked by those identities publicly. For example, an individual who lives in an area that treats having sex with someone of the same gender identity as a felony crime may be more willing to accurately identify their sexual orientation if their response is not stored in a central database (that could be rendered to a government agency) and cannot be de-anonymized.

RECOMMENDATION(S)

- If organizations are employing data inference techniques for demographic characteristics, organizations should provide individuals with complementary opportunities to self-identify or to check their ascribed demographics.
- Teams should account for sampling bias by doing specific outreach to communities at risk of underrepresentation.
- Organizations should ensure participants are provided with clear, accessible opportunities to accept or refuse participation with an informed understanding of the privacy protection provided to them, what their data will be used for, and for how long their data will be retained.

Design Considerations for Differentially Private Federated Statistics

Differentially private federated statistics is an approach with many different features. How these features are defined can impact whether differentially private federated statistics can strengthen or impede an algorithmic fairness strategy. The following section reviews the different aspects of differentially private federated statistics to consider when designing a robust algorithmic bias assessment strategy.

The Differential Privacy Model

Differentially private federated statistics can be designed with either local differential privacy and central (or global) differential privacy. Each of these models imposes privacy guarantees at different levels, and therefore have implications for individual privacy and the amount of trust that an individual is required to have in the organization conducting the fairness analysis.

In a local differential privacy model (LDP), statistical noise is added to an individual's data before it is shared from their device with the central server. As a result, no raw data is shared with the organization conducting the analysis. This LDP model (Image 1) addresses the privacy concerns held by individuals when considering whether or not to contribute their data. It removes the need for individuals to trust the organization. Alongside the LDP model, organizations also have an option to incorporate a secure aggregation protocol. This is an additional privacy guarantee that is enforced after data (with statistical noise) has been aggregated from multiple individual devices on the central (or third-party) server to ensure that aggregate data will not be released unless the aggregate privacy guarantee is achieved.

In contrast, a central differential privacy (CDP) model adds statistical noise at the aggregate level once received by the central server. In this model, the central aggregator has access to an individual's raw data before adding noise to achieve the differential privacy guarantee (Image 2). The disadvantage to this model is that it requires individuals to have more trust in the organization to protect their privacy, and therefore can negatively impact an individual's willingness to contribute their data, leaving the organization with a less comprehensive data pool for their analysis.

IMAGE 1: Local Differential Privacy

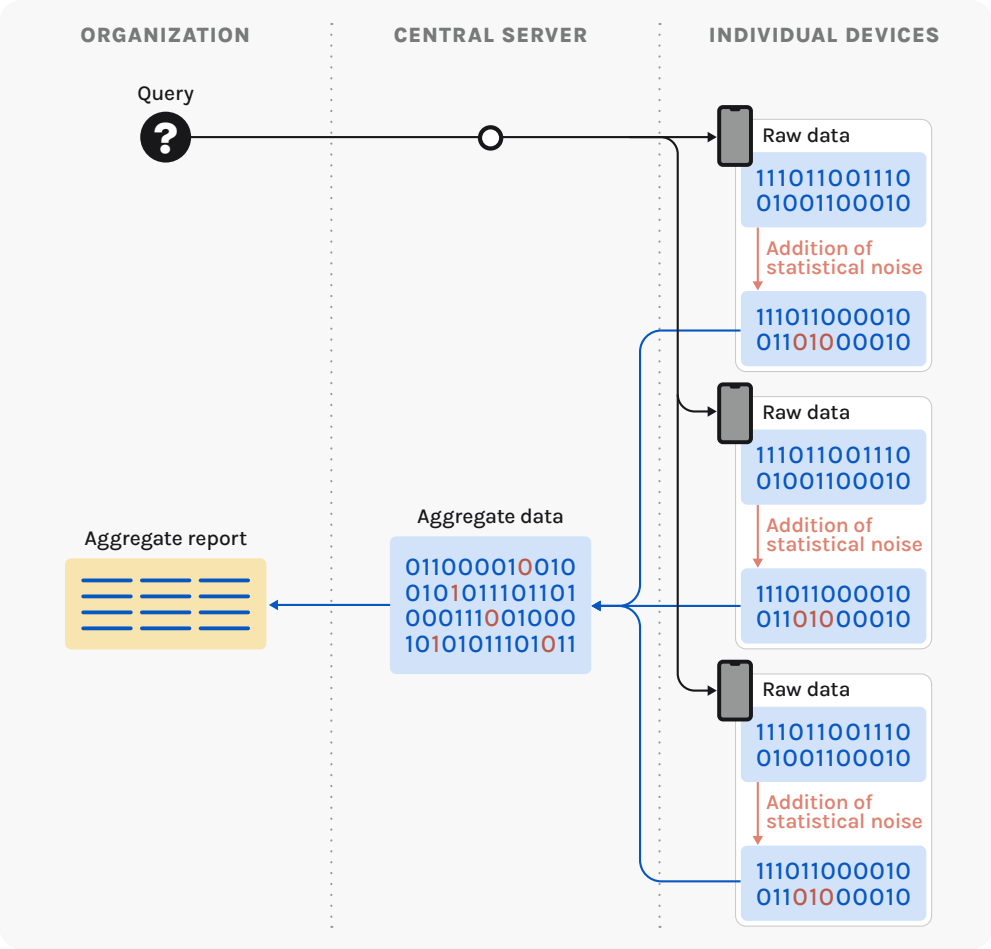
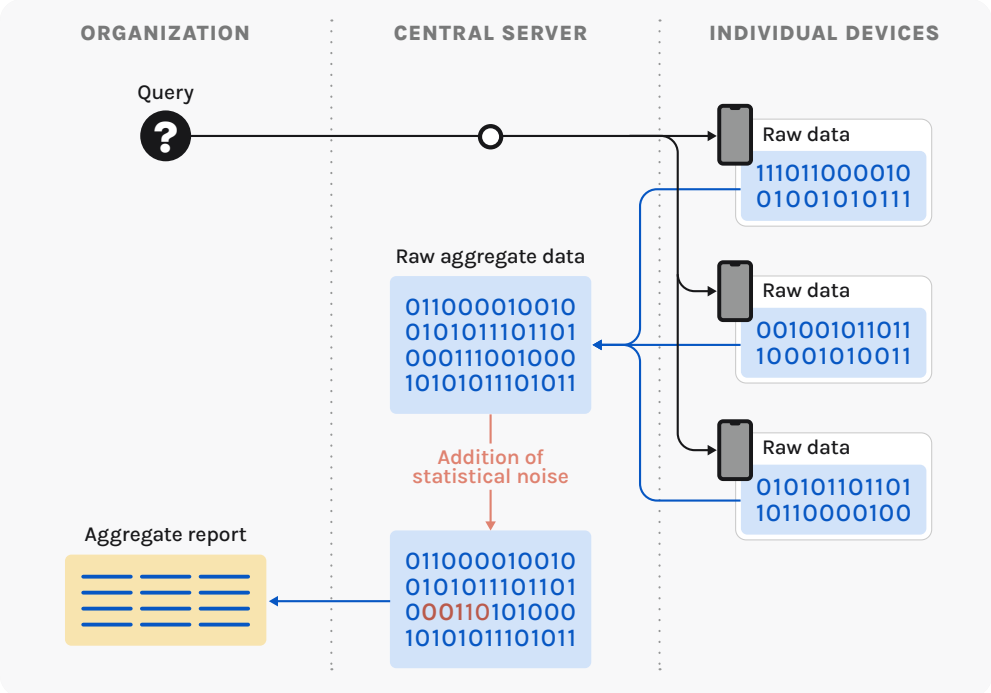


IMAGE 2: Central Differential Privacy



RECOMMENDATION(S)

- Organizations should use local differential privacy (LDP) with secure aggregation to guarantee the highest amount of privacy protection for individuals who share their data.
- Organizations should consider incorporating a secure aggregation protocol alongside LDP to bolster privacy once data is received by the central server.

The Privacy Budget

The privacy budget, or epsilon, is a commonly accepted metric of privacy loss. The epsilon value can be seen as a spectrum, with absolute data privacy on one end and absolute data accuracy on the other: the smaller the epsilon value, the greater the privacy guarantee is for individuals, and the less utility (or accuracy) the aggregated data has. An epsilon value of 0 or 1 is considered highly private as, while a value between 2 and 10 is considered as providing some privacy, whereas a value above 10 is considered to provide little to no privacy since very little noise will be added to the original data to prevent an adversary from recognizing that a revealing output has occurred. An epsilon value can be set at the local or aggregate level under differentially private federated statistics.

When designed with strong privacy guarantees, differentially private federated statistics has the ability to ensure an individual's sensitive data remains private and their identity remains unidentifiable in aggregate. However, while a smaller epsilon value will yield greater privacy, data analysis outputs will be less accurate (relative to the same analysis conducted with a higher epsilon value). In some instances, it may be necessary to set a higher epsilon value as more accurate data analysis is needed (e.g., analyses of groups with lower population sizes).

It is important to note that the application of differentially private federated statistics is not in and of itself a privacy guarantee: a deployment of differential privacy is only as private as the choice of epsilon. For example, organizations using differential privacy may overstate the extent of data privacy they are providing if they use a high epsilon value. By using differentially private federated statistics with a small(er) epsilon value, organizations can neutralize linkage attacks, or the ability to identify an individual using data from multiple datasets to establish a link and reveal their identity.

As noted earlier, enforcing a small epsilon — and therefore a higher degree of privacy — does pose tradeoffs for accuracy (i.e., the ability for an individual to be accounted for in a fairness assessment), particularly for demographic groups that make up statistical minorities in datasets as is often the case for marginalized communities. A smaller epsilon can render statistical minority groups invisible during analysis. This could lead to inaccurate, ineffective, or even harmful fairness assessments if small populations continue to be excluded from fairness assessments. Adding to the complexity, a larger

epsilon reduces privacy guarantees which can be particularly harmful for marginalized communities who are already at higher risk of surveillance.

There is no universal balance standard for setting the epsilon. This decision depends on multiple factors including the risks to individuals and communities associated with data collection, levels of desired privacy, organizational needs, operational definition of fairness and other context specific to the fairness assessment. In some instances, it may be necessary — and even actively supported by the population of users — to set a higher epsilon value so assessment of their small group population may be more accurate and potentially subject to remediation if bias is identified.

RECOMMENDATION(S)

- Teams should choose the epsilon and other privacy parameters with the needs of those most at risk of algorithmic harm as the priority for analysis and investigation.
    - Focusing attention on the study of those most at risk, even when they make up a statistical minority, can generate benefits for all users.

Queries

Queries refer to pre-defined and approved functions (i.e. inquiries) that can be applied to raw data on devices resulting in an aggregate report. Choosing the correct query parameters is crucial to designing an effective fairness assessment process as the queries determine what instances of bias are or are not visible to the data analysts. For example, in assessing the functionality of a computer vision algorithm for any potential bias, a relevant query may be related to performance across gender groups for a racially diverse population given known issues of computer vision algorithms and intersectional gender bias. The chosen definition of fairness can help organizations in defining query parameters. Participatory methods, such as the inclusion of impacted communities and interdisciplinary experts, may also be employed to help identify the initial set of queries to be used to explore algorithmic bias.

When designing queries, organizations might consider factors such as the query frequency, amount permitted, and content. Within the differentially private federated statistics model, data analysts can deploy queries adaptively and tailor their queries based on previously observed responses. This poses some necessary considerations for how queries are determined and overseen. It is possible that individual devices could be "mined" excessively by deploying an endless number of queries. Adaptive queries could also make it possible to trace data points to specific devices by tailoring the questions in an increasingly more specific way to discern the device (and therefore the user and the user's data) from another.

The practice of data minimization ensures organizations only collect the data necessary to accomplish a given task. This can be achieved through a thoughtful query design and

approval process in which only necessary information is retrieved from individual's devices and aggregated in reports. Additionally, with the appropriate restrictions and monitoring in place, it is possible to employ differentially private federated statistics to minimize the risk of, for example, re-identification using the individual data that is collected and used. For example, the differential privacy constraint can be applied to sets of queries. Instead of saying "every query must satisfy *X privacy constraint*," the rule can state "the set of all queries asked this month must satisfy *X privacy constraint*."[38]

38 "Privacy constraint" refers to the set of rules (which assign privacy levels) to the dataset being analyzed.

Combined with retention limits, this can help effectively limit privacy loss. This can also help organizations stay in compliance with data regulation standards such as GDPR, CCPA, and HIPPA. Data minimization also helps to mitigate social harms stemming from broad data collection, such as increased surveillance and data misuse or use beyond informed consent.

Ultimately, it is important for organizations to balance varying constraints such as the need for data minimization as well as the need to obtain adequate levels of responses to conduct a robust fairness assessment.

RECOMMENDATION(S)

- Teams should ensure query parameters align with the definition of fairness.
- Teams should work with interdisciplinary experts and/or community groups in designing query parameters.
- Teams should balance data minimization with the need for robust fairness assessment depending on specific context.

Data Retention

Data retention, which refers to the length of time individual data can be accessed via a query or used in aggregate before being destroyed, is an important consideration due to real and perceived threats of data breaches, potential re-identification, and being repeatedly queried. Differentially private federated statistics on its own does not prohibit the long-term storage of data, whether on the device or in a central server. This is a determination an algorithmic assessment team or the organization must establish for itself. However, this is not a straightforward choice, especially when considering the needs of marginalized communities with small populations.

A short data retention period (e.g., 30 days), for example, protects an individual from being queried many times, but small groups may be underrepresented in the overall data population at any given time. For example, if there are only 20 people out of 500 who identify as a member of a small religious sect, all 20 people would need to consent to their data usage within the same period of time in order to exceed the privacy budget threshold to be accurately analyzed as a group.

A long data retention period (e.g., two years), on the other hand, may subject an individual device to participating in many assessments, perhaps exceeding the expectations of the individual who consented to use of their data. Such a situation would also make it more likely an individual device could be re-identified if it appears across many query reports.

RECOMMENDATION(S)

- Organizations should institute a data retention period to ensure individual data is not perpetually used or accessible.
- Organizations should think carefully about how the data retention period will impact their ability to identify bias when users are able to contribute their data across a long time period, particularly for statistical minorities who may not all contribute their data at once.

Conclusion

An important part of responsible AI development is recognizing that it is difficult, if not impossible, to release an algorithmically-driven feature or product that is guaranteed to work every time for all people and situations. Rigorous pre- and post-deployment fairness assessments are necessary to surface any potential bias in algorithmic systems. Post-deployment fairness assessments can pose additional challenges to organizations, as they often involve collecting new user data, including sensitive demographic data, to observe whether the algorithm is operating in ways that disadvantage any specific group of people. The collection and use of demographic data is recognized to be challenging for organizations due to concerns related to data privacy, data security, and legal barriers. Demographic data collection also poses key risks to data subjects and communities such as data misuse or abuse of data (including potential discriminatory uses), as well as harms stemming from misrepresentation and miscategorization in datasets.

In an effort to deploy algorithmically driven features responsibly, Apple introduced IDs in Apple Wallet with mechanisms in place for Apple (and the identification card issuing state authority) to identify any potential biases users may experience when setting up or using their new digital ID. Currently only available in the United States, Apple applied differentially private federated statistics as a way to protect users' data, including their demographic data, as part of IDs in Apple Wallet. The main benefit of using differentially private federated statistics is the preservation of data privacy by combining the features of differential privacy (e.g., adding statistical noise to data to prevent re-identification) and federated statistics (e.g., analyzing user data on individual devices, rather than on a central server, to avoid the creation of datasets that can be hacked or otherwise misused).

A member organization of Partnership on AI (PAI), Apple shared details about the use of differentially private federated statistics in a US context for discussion by responsible AI, algorithmic fairness, and social inequality experts across two convenings. Independently organized and hosted by PAI, the two expert convenings discussed how algorithmic fairness assessments are strengthened, challenged, or otherwise unaffected by the use of differentially private federated statistics. PAI applies a sociotechnical lens to various AI issues, including algorithmic fairness and bias issues, in order to draw attention to the complex ways AI can have social impact, particularly for marginalized demographic groups.

Expert participants were asked to consider not only the specific technical strengths or weaknesses of differentially private federated statistics but how this approach interacts with an overall algorithmic fairness strategy. Recognizing that data privacy and security are not the only concerns people have regarding the collection and use of their demographic data, participants were directed to consider whether differentially private federated statistics could also be leveraged to attend to some of the other social risks that can arise.

The expert participants — drawn from commercial AI companies, research institutions, and civil society organizations — emphasized the importance of having both pre- and post-deployment algorithmic fairness assessments throughout the development and deployment of an AI-driven system or product/feature. Post-deployment assessments are especially important as they enable organizations to monitor algorithmic systems once deployed in real-life social, political, and economic contexts. They also recognized the importance of thoughtfully collecting some demographic data in order to help identify group-level algorithmic harms.

The expert participants, however, clearly noted that a secure and privacy-preserving way of collecting and analyzing sensitive user data is, on its own, insufficient to deal with the risks and harms of algorithmic bias. In fact, they expressed that such a technique is not entirely sufficient for dealing with the risks and harms of collecting demographic data. Instead, the convening participants identified key choice points facing AI-developing organizations to ensure the use of differentially private federated statistics contributes to overall alignment with responsible AI principles and ethical demographic data collection and use.

The following tables (Tables 2 and 3) summarize the different choice points and recommendations for best practices identified by the expert participants. Recommendations are organized into two types:

1. general considerations that any AI-developing organization should consider for their post-deployment algorithmic fairness assessment (Table 2)
2. design choices specifically related to the use of differentially private federated statistics within a post-deployment algorithmic fairness strategy (Table 3)

The choice points identified by the expert participants summarized in Table 2 emphasize the importance of carefully applying differentially private federated statistics in the context of algorithmic bias assessment. They noted that several features of the technique can be determined in such a way that reduces the efficacy of the privacy-preserving and security-enhancing aspects of differentially private federated statistics. Several expert participants highlighted Apple's decision to limit the data retention period (90 days), clearly and simply sharing what data the user will be providing for the assessment, and maintaining organizational oversight of the query process and parameters as aligning with the best practices they would recommend.

Many of the recommendations surfaced by the expert participants focus on the resources (e.g., financial, time allocation, and staffing) necessary to achieve a level of alignment and clarity on the nature of "fairness" and "equity" AI-developing organizations are seeking for their AI-driven tools and products/features before integrating differentially private federated statistics into their overall bias mitigation strategy. While these considerations may seem tangential, the experts emphasized the importance of establishing a robust foundation on which differentially private federated statistics could be effectively utilized. Any form of demographic data collection or use can expose people to potential risk or harm.

Regardless of the steps taken to minimize such risk, the collection of demographic data without an explicit purpose or effective plan for its responsible usage is not justifiable given the potential individual or societal cost. Differentially private federated statistics, in and of itself, does not mitigate all the potential risks and harms related to collecting and analyzing sensitive demographic data. It can, however, strengthen overall algorithmic fairness assessment strategies by supporting better data privacy and security throughout the assessment process.

TABLE 2: General Considerations for Algorithmic Fairness Assessment Strategies

Choice Point	Recommendation(s)
Establishing organizational support	• Organizations should provide teams with adequate time and resources to design and deploy algorithmic fairness assessments. • Teams should obtain executive, leadership, and middle management buy-in to ensure they receive the proper support to effectively address any bias identified. • Team members involved in conducting the overarching fairness assessment, which differentially private federated statistics is one component of, should ensure meaningful engagement with non-technical experts and community groups to inform their overall approach. • This involves setting expectations, maintaining communication, and providing compensation for those external to the organization who contribute their time and expertise.
Defining Fairness	• Organizations should achieve alignment between technical and non-technical definitions of fairness. • Organizations should achieve alignment between developer and user or public understanding and measurement of fairness. • Organizations must practice transparency when it comes to how they define fairness.
Identifying relevant demographic categories	• Organizations should allocate necessary resources to conduct original research into what demographic categories are relevant as well as how communities that interact with the algorithmic system define themselves to yield a more inclusive fairness assessment process.
Determining the data collection method(s)	• Organizations should provide individuals with complimentary opportunities to self-identify or to check their ascribed demographics if using inference techniques. • Teams should account for sampling bias by doing specific outreach to communities at risk of underrepresentation. • Organizations should ensure participants are provided with a clear, accessible opportunity to accept or refuse participation with an informed understanding of the privacy protection provided to them, what their data will be used for, and for how long their data will be retained.

TABLE 3: Design Considerations for Differential Private Federated Statistics

Choice Point	Recommendation(s)
Choosing the differential privacy model (local differential privacy vs. central differential privacy)	• Organizations should use local differential privacy (LDP) to guarantee the highest amount of privacy protection for individuals who share their data. • Organizations should consider incorporating a secure aggregation protocol alongside LDP to bolster privacy once data is received by the central server.
Determining the appropriate privacy budget/epsilon	• Teams should choose the epsilon and other privacy parameters with the needs of those most at risk of algorithmic harm at the center (which will often benefit all users) rather than choosing based on the needs of the majority of users as this could exacerbate existing inequities.
Designing queries	• Teams should ensure query parameters align with the definition of fairness. • Teams should work with interdisciplinary experts and/or community groups in designing query parameters. • Teams should balance data minimization with the need for robust fairness assessment depending on specific context.
Determining the data retention period	• Organizations should institute a data retention period to ensure individual data is not perpetually used or accessible. • Organizations should think carefully about how the data retention period will impact their ability to identify bias when users are able to contribute their data across a long time period, particularly for statistical minorities who may not all contribute their data at once.

Acknowledgments

We would like to thank the experts who participated in our multistakeholder convenings whose contributions were invaluable to the development of this paper. They include Albert Fox Cahn, Esq. (Executive Director, Surveillance Technology Oversight Project), Aleksandra (Sesa) Slavkovic, Ph.D. (Professor of Statistics and Public Health Sciences, Huck Chair in Data Privacy and Confidentiality, Penn State University), Aloni Cohen (Assistant Professor, University of Chicago), Alycia N. Carey (Distinguished Doctoral Fellow, University of Arkansas), Amina Abdu (University of Michigan), Andrew Smart (Researcher, Google Researcher), Arjun Subramonian (Ph.D. Student, University of California, Los Angeles), Berk Ustun, Ph.D (Assistant Professor, UCSD), Brandie Nonnecke, Ph.D. (Director, CITRIS Policy Lab & Assoc. Research Prof., Goldman School of Public Policy, UC Berkeley), Caroline Siegel Singh (Program Manager, The Greenlining Institute), Dr Kerry McInerney (Leverhulme Centre for the Future of Intelligence), Dr Louise Hickman (Research Associate, The Minderoo Centre for Technology and Democracy, University of Cambridge), Dr. Muneeb Ul Hassan (Associate Lecturer in Cybersecurity, Deakin University, Australia), Ferdinando Fioretto (Assistant Professor, University of Virginia), Julie M. Wenah, Esq. (Digital Civil Rights Coalition and Women In Product Board Director), Londa Schiebinger, Ph.D. (John L. Hinds Professor of History of Science, Stanford University), Lydia X. Z. Brown, J.D. (Lecturer in Disability Studies and Women's and Gender Studies, Georgetown University), Matt Canute (Digital Democracies Institute at Simon Fraser University), Micah Altman (Research Scientist, Center for Research in Equitable and Open Scholarship, MIT), Miranda Bogen, Ninareh Mehrabi, Ph.D. (Scientist, Amazon), Orestis Papakyriakopoulos (Sony AI / TU Munich), Priyanka Nanayakkara (PhD Candidate, Northwestern University), Rachel Cummings, Ph.D. (Associate Professor, Columbia University), Reva Schwartz (Research Scientist, National Institute of Standards and Technology (NIST)), Sikha Pentyala (Research Assistant, University of Washington Tacoma), Sikha Pentyala (Research Assistant, University of Washington Tacoma), Suresh Venkatasubramanian (Professor of Computer Science and Data Science, Brown University), Wells Lucas Santo (University of Michigan), Xintao Wu, Ph.D. (Professor, University of Arkansas).

We would also like to thank our PAI colleagues who made this publication possible, including Albert Tanjaya, Hudson Hongo, Jason Millar, Neil Uhl, Rebecca Finlay, and Stephanie Bell.

Funding Disclosure

This study was funded entirely by Partnership on AI, including all costs associated with hosting two virtual convenings. Partnership on AI is funded through a combination of philanthropic institutions and corporate charitable contributions. As a founding Partner, Apple provides Partnership on AI with corporate charitable contributions. However, primary corporate funding is always considered general operating support and legally classified as non-earmarked charitable contributions (not donations in exchange for goods or services, or quid pro quo contributions) to avoid the possibility of conflict in corporate funders having undue influence on Partnership on AI's agenda, initiatives, or any specific projects. At the time of initial publication, an employee of Apple, Inc. served on the board of directors of Partnership on AI. More detail on Partnership on AI's funding and governance is available online.

Appendices

APPENDIX 1
Fairness, Transparency & Accountability Program Area at Partnership on AI

The Fairness, Transparency, and Accountability program area at Partnership on AI encompasses PAI's large body of research and programming around issues related to discriminatory harms of algorithmic systems. Since 2020, the team has sought to understand the types of demographic data collection practices and governance frameworks required to ensure that fairness assessments of algorithmic systems are conducted in the public interest. The team has explored data collection and algorithmic fairness practices and processes from both an organizational process perspective and an equity and inclusion perspective. The program area aims to demonstrate the importance of categorization and datafication practices to organizational efforts to: 1) make algorithmic decision-making more "fair"; 2) develop guidelines for how organizations can include participatory, inclusive practices around data collection to achieve "fairness" or "non-discrimination"; and 3) assess the contextual feasibility of existing and emerging fairness techniques.

APPENDIX 2
Case Study Details

In 2022, Apple released a new feature in their Apple Wallet app, IDs in Wallet, which allows users to store a digital copy of their state-issued identification card or driver's license to be used in lieu of their physical card. Users are required to undergo several identification verification checks to help ensure that the person adding the identity card to Wallet is the same person to whom the identity card belongs. The state is responsible for verifying and approving the user's request to add their driver's license or state ID to Wallet.

In order to help Apple and the state issuing authority ensure fairness in the identity verification process, Apple asks users to share select demographic data (such as age range or sex) from a user opt-in screen at the end of the ID in Wallet setup flow. This analysis helps determine if outcomes during the setup and approval process are different for groups of users.

Sharing information is optional and if users agree to share, the information is collected in a way that helps preserve user privacy. Federated statistics uses differential privacy to allow analytics of aggregated information without Apple, or the ID-issuing state, learning individual-level information. No personally identifiable information is collected, stored, or used by Apple or the state issuing authority as part of this process. Users can opt out of sharing this data at any time.

Apple is using differentially private federated statistics as part of a larger fairness assessment strategy for IDs in Wallet that includes rigorous user testing and inclusivity roundtable discussions with state issuing authorities and third-party vendors, and integration of existing inclusivity strategies from other Apple teams, among others.

APPENDIX 3
Multistakeholder Convenings

PAI led a series of convenings designed to engage a diverse set of experiences and expertise to explore the questions posed in this project through both social and technical lenses. PAI has had success with facilitating multistakeholder, multi-disciplinary discussions about pressing ethical issues in the field of AI, leveraging the diversity of participants to capture the necessary nuance to address those issues.

Extended semi-structured small group discussions moderated by facilitators from PAI allowed expert participants to uncover questions and considerations around the use of differential private federated statistics for the algorithmic fairness assessments that may not have been considered otherwise.

PAI actively encourages participants to pose additional questions or considerations outside of the initial interview protocol. By grounding these convenings in social scientific research methodologies, discussions are designed to yield more general insights on how organizations can use differentially private federated statistics to ethically and responsibly approach bias and fairness assessments, or to more rigorously examine and improve their own approaches and strategies.

APPENDIX TABLE 1: Convening Discussion Questions

Convening Topic	Discussion Questions
Methods for Inclusivity in Data Collection	• When beginning a data collection effort to support algorithmic fairness, what processes should be adopted to support the identification of the appropriate and most accurate measurement of demographic categories to include in a fairness assessment? • What considerations should be taken into account when deciding on the desired demographic data? • Who should be included in this process? • What impact might these considerations have on different demographic groups? • How can we design participatory and inclusive demographic data collection methods that preserve privacy and advance fairness? • What are the benefits and risks of various methods (perhaps those we discussed previously), particularly for marginalized groups? • What role does consent play in this process?
Advancing Privacy and Fairness	• Does this technique as a whole lend itself to both privacy and accuracy for users from all demographic groups? • Which components of differentially private federated statistics are most vulnerable to breakdowns that could harm marginalized groups? • How should interacting components of differential private federated statistics be determined when striving towards privacy and accuracy for all demographic groups? • Under what circumstances would the privacy budget, sets of queries, and lifespan of data retention be adjusted? • Could this be adjusted in order to better serve different demographic groups?

APPENDIX 4
Glossary

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

For example, an algorithmic system that uses racial proxy analysis by assigning a racial category to an individual based on the analysis of an individual's skin tone in an image may classify someone as "White" due to the perceived "light" skin tone of an individual who racially identifies as "Asian."

Algorithmic system
Refers to a system composed of one or more algorithms, or an automated procedure used to perform a computation, and includes systems using machine learning or following a pre-programmed set of rules.

For example, search engines, traffic signals, and facial recognition software all rely on algorithmic systems to function.

Analytical accuracy
Refers to the closeness (accuracy) between the representation of a value or data point and the true value or data point.

For example, high analytical accuracy is achieved when the distribution of gender categories (% of a population in each gender category) in a dataset matches the gender distribution of the measured population.

Data breach
Refers to an incident where sensitive data or confidential information is stolen or otherwise accessed without the authorization of the system's owner.

For example, the physical theft of a hard drive containing users' personal information or a ransomware cyberattack that prevents a company from accessing its own customer data unless a ransom is paid are both considered data breaches.

Datafication
Refers to the conversion of various aspects of human life into quantitative data which then allows for quantitative analysis of social and individual behaviors.

For example, peoples' communication, images, and speech are all converted into data points via various technological platforms by messaging platforms (e.g., text from conversations), social media sites (e.g., text- and image-based posts), and digital assistants (e.g., audio recordings of voice commands given to the device), respectively.

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.

For example, if a survey only provides the options to select "man" or "woman," an individual who identifies as gender non-binary (as neither a woman or a man) will be represented inaccurately in the dataset.

Data re-identification or de-anonymization
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, typically by matching the anonymized dataset with publicly available data or auxiliary data.

For example, an anonymized dataset containing private health information can be re-identified if the identification number used to distinguish individuals from one another is their social security number and a separate list with the individuals' names and social security number is used.

Demographic inference
Refers to a set of techniques used by data analysts to fill in, as accurately as possible, missing or unidentified demographic traits by analyzing other available data.

For example, an individual's name can be used to guess their gender identity by analyzing how often that name is associated with individuals who identify as a woman versus how often it is associated with individuals who identify as a man.

Disparate impact
(and adverse impact and the 80% rule)
Refers to the often-used legal interpretation of "fairness" where the emphasis is on determining whether one group experiences different outcomes or treatment (unfair), including when differences emerge unintentionally.

Adverse impact specifically refers to instances of disparate impact when a group disproportionately (with greater frequency or intensity) experiences a negative outcome or treatment.

The "80% rule" refers to one specific, and simple, way to "test" for instances of adverse impact, originally designed by the State of California Fair Employment Practice Commission, by calculating whether the selection rate for a minority group (the group with the lowest selection rate) is less than 80% of the rate for the group with the highest selection rate (typically the majority group).

Federated statistics or federated learning

Refers to an approach of user data analysis that does not transfer data from individual devices to a central server for analysis, but instead runs local computation on the individual devices and sends the results to a central server where composite results are made available. Federated learning most often refers to an approach used to train machine learning models on individual devices, while federated statistics most often refers to basic statistical analysis conducted on individual devices.

Knowledgeable consent

Refers to a form of consent (permission for something to happen) where individuals fully understand the scope and implications of their participation (including disclosure of their information) and have the ability to refuse or withdraw their participation at any time.

For example, a person is able to provide knowledgeable consent for the use of their blood sample in a scientific experiment because the individual understands the various instances when their sample may be used — and what may happen due to the use of their sample — and is given frequent opportunities throughout the entire duration (when their sample is in possession of the scientists conducting the experiment) to withdraw their sample from use.

Model poisoning attack

Refers to a vulnerability of federated learning (due its distributed nature) where the model (or data analysis) being trained through federated learning is attacked through the insertion of "bad" (inaccurate or irrelevant) data into an AI model's training dataset, leading the algorithmic model to complete unintended learning and creating undesirable variability in the model's outputs.

Secure aggregation

Refers to a cryptographic protocol (a system of rules related to how should be structured and represented and how algorithms should be used specifically to prevent third parties from accessing or revealing information) that securely computes the aggregation (analysis or compilation) of its inputs in such a way that the inputs are kept hidden and the central server where the aggregation takes place or is stored cannot learn the value of the individual inputs.

Secure multi-party computation

Refers to a technique designed to cryptographically protect information by allowing multiple, distinct parties — each holding their own private data — to conduct a collaborative computation (combine and analyze information together) without revealing the specific details or value of that data to the other parties, thereby protecting the privacy of the data while not endangering the security of the data.

Sociotechnical

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

For example, a sociotechnical analysis of an algorithmic system would include assessment of the various social components influencing the design, production, and deployment of the system as well as the social impacts of the system.

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 on the basis of a demographic trait.

For example, people who identify as transgender (someone whose gender identity differs from the one that is typically associated with the sex assigned at birth) is a social minority in the US due to the discrimination and harm experienced and a statistical minority due to the relatively smaller size of the population compared to people who identify as cisgender (someone whose gender identity corresponds with the sex assigned at birth). On the other hand, women are considered a social minority due to the systematic discrimination the group experiences but are not currently a statistical minority.

Statistical 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.

For example, statistical parity in a machine learning model used to recommend promotions at a workplace would require that men and women in the dataset have the same likelihood of receiving a recommendation for promotion.

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.

Surveillance infrastructure can include tools like traffic light cameras or wearable devices. Due to systematic inequalities like racism and xenophobia which consider specific groups of people as more dangerous or socially deviant (and therefore require constant monitoring), marginalized communities are often disproportionately subject to surveillance infrastructure.

APPENDIX 5

Detailed Summary of Challenges & Risks Associated with Demographic Data Collection & Analysis

APPENDIX TABLE 2: Challenges and Risks Associated With Demographic Data Collection and Analysis

Challenge or Risk	Definition	Example
ORGANIZATIONAL CONCERNS		
Organizational priorities	Fairness analyses and interventions often do not support, or may even conflict, with key performance indicators used to evaluate employee performance	Adequate (or additional) data collection may be considered too costly to be financially justifiable for an organization whose primary concern is to reduce development and production costs as much as possible to maximize overall profitability private-sector company
Public relations risk	Efforts to collect demographic data could lead to public suspicion and distrust	Due to increasing public scrutiny of data misuse by organizations, the public is skeptical of any reason given to justify the collection of additional user data and looks for any indication (whether real or imagined) of data misuse
Discomfort (or lack of expertise) with identifying appropriate demographic groups	The lack of standardized approaches to choosing salient demographic categories and subcategories leads to inaction	Companies are hesitant to define demographic categories in collection efforts at risk of public criticism so often turn to outdated governmental standards, like binary gender categories
LEGAL BARRIERS		
Anti-discrimination laws	In key protected domains, such as finance and healthcare, collection of demographic data may conflict with anti-discrimination laws	Companies selling credit-based products, for example, are barred from collecting demographic data in most instances but are still held to anti-discrimination standards, making robust fairness assessments difficult
Privacy policies	Increasingly protective privacy regulation has empowered privacy and legal teams to err on the side of caution when it comes to data sensitivity	The GDPR designation of race as a "special" demographic category that requires companies meet a high set of standards in order to justify collection may dissuade companies from gathering this information
SOCIAL RISKS TO INDIVIDUALS		
Unique privacy risks associated with the sharing of sensitive attributes likely to be the target of fairness analysis	As attributes such as race, ethnicity, country of birth, gender, and sexuality are usually consequential aspects of one's identity, collecting and usage of 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

Possible harms stemming from miscategorizing and misrepresenting 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 which assume invisible aspects of one's identity from visible characteristics, such as physiognomy
Use of sensitive data beyond data subjects' expectations	Use of demographic information beyond initial intent not only breaks consent of data subjects but can also lead to unintended harmful consequences	The US government developed the Prisoner Assessment Tool Targeting Estimated Risk and Needs to provide guidance on recidivism reduction programming but was then repurposed to inform inmate transfers, leading to racially disparate outcomes
SOCIAL RISKS 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 can be further exacerbated by fairness assessments. Expanded surveillance can constrain agency and result in exploitation for these groups.	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, leading to undercounting of gender non-conforming individuals
Data subjects ceding the ability to define for themselves what constitutes biased or unfair treatment	When companies leading the data collection effort alone define unfairness, with no input from marginalized groups, key instances of discrimination can be missed and the status quo can be reinforced	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