The Privacy-Bias Tradeoff: Data Minimization and Racial Disparity Assessments in U.S. Government

Arushi Gupta
Stanford University
Stanford, California, USA

Victor Y. Wu
Stanford University
Stanford, California, USA

Helen Webley-Brown
Massachusetts Institute of Technology
Boston, Massachusetts, USA

Jennifer King*
Stanford University
Stanford, California, USA

Daniel E. Ho*
Stanford Law School
Stanford, California, USA

ABSTRACT

An emerging concern in algorithmic fairness is the tension with privacy interests. Data minimization can restrict access to protected attributes, such as race and ethnicity, for bias assessment and mitigation. Less recognized is that for nearly 50 years, the federal government has been engaged in a large-scale experiment in data minimization, limiting (a) data sharing across federal agencies under the Privacy Act of 1974, and (b) data collection under the Paperwork Reduction Act. We document how this "privacy-bias tradeoff" has become an important battleground for fairness assessments in the U.S. government and provides rich lessons for resolving these tradeoffs. President Biden's 2021 racial justice Executive Order 13,985 mandated that federal agencies conduct equity impact assessments (e.g., for racial disparities) of federal programs. We conduct a comprehensive assessment across high-volume claims agencies that affect many individuals, as well as all agencies filing "equity action plans," with three findings. First, there is broad agreement in principle that equity impact assessments are important, with few parties raising privacy challenges in theory and many agencies proposing substantial efforts. Second, in practice, major agencies do not collect and may be affirmatively prohibited under the Privacy Act from linking demographic information. This has led to pathological results: until 2022, for instance, the US Dept. of Agriculture imputed race by "visual observation" when race information was not collected. Data minimization has meant that even where agencies want to acquire demographic information in principle, the legal, data infrastructure, and bureaucratic hurdles are severe. Third, we derive policy implications to address these barriers.

ACM Reference Format:
Arushi Gupta, Victor Y. Wu, Helen Webley-Brown, Jennifer King, and Daniel E. Ho. 2023. The Privacy-Bias Tradeoff: Data Minimization and Racial Disparity Assessments in U.S. Government. In *2023 ACM Conference on Fairness, Accountability, and Transparency (FAccT '23), June 12–15, 2023, Chicago, IL, USA.* ACM, New York, NY, USA, 17 pages. https://doi.org/10.1145/3593013.3594015

*Equal co-supervision

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

ACM ISBN 979-8-4007-0192-4/23/06.
https://doi.org/10.1145/3593013.3594015

1 INTRODUCTION

While early work in algorithmic fairness has grappled with the fairness-accuracy tradeoff [32, 52, 70, 73, 120, 148], one of the emerging tradeoffs is between fairness and individual privacy [25, 67]. Much technical work, for instance, has shown the tension between applying differential privacy and achieving algorithmic fairness [51]. Differential privacy [40] can worsen disparate impact in model accuracy [13], such as for racial minorities or rural communities [27, 125, 144]; equalized odds as a fairness measure can disproportionately leak information for disadvantaged groups [25]; and the impossibility theorems that have vexed algorithmic fairness have similarly affected simultaneously achieving differential privacy and fairness [6]. Institutionally, leading work has shown how interpretations of privacy law and policy, such as the E.U.'s General Data Protection Regulation, have undermined the capacity to assess and mitigate bias in the private sector, as technical teams are not allowed to access protected attribute information such as race and ethnicity [9, 146]. Data minimization – the principle that entities should collect and retain only data minimally necessary to achieve their objectives – has meant that critical information needed to conduct fairness assessments is unavailable. We call this the emerging "privacy-bias tradeoff." As companies and regulators turn toward protecting individuals' information privacy via data minimization,[1] we ask: How can we ensure that the lessons of algorithmic fairness are not ignored? How have institutions attempted to grapple with these tensions? And what practical policy options are available to navigate the tradeoff in the most grounded fashion?

What is less well-known is that since 1974, the U.S. government can be described as engaging in a large-scale data minimization experiment. The Privacy Act of 1974 mandates that federal agencies (a) collect personally identifiable information only as necessary to execute their statutory mandate, (b) use this information only for the purpose that justified its collection, and (c) refrain from data sharing or linkage [77]. Statutes like the Paperwork Reduction Act of 1980 make it procedurally challenging for agencies to add new mechanisms for data collection (e.g., surveys, web forms, paper forms, or revisions thereof), typically requiring approval by the Office of Management and Budget (OMB) and a notice-and-comment process for public input when they do try to do so.

[1] The E.U.'s General Data Protection Regulation incorporated data minimization into its framework with its adoption in 2018. In 2022, the U.S. Federal Trade Commission signaled that it may be considering adopting data minimization provisions to regulate the private sector during its Advanced Notice of Proposed Rulemaking [29].

In this paper, we demonstrate that the "privacy-bias tradeoff" has become an important battleground for fairness assessments in U.S. government. On his first day in office, President Biden signed Executive Order (EO) 13,985, a racial justice initiative mandating that agencies conduct "equity assessments" of federal programs [48]. For the first time, agencies are required to assess disparities (e.g., along race and ethnicity) in accessing benefits and opportunities in federal policies and programs. Such assessments have been uniquely important for understanding disparities with the rise of algorithmic decision-making tools, as emphasized in the nondiscrimination principle of the Blueprint for an AI Bill of Rights [98], the Trustworthy AI Executive Order 13,960 [46], and the second racial justice Executive Order 14,091 [47]. Yet in practice, as we show, the Privacy Act has made the implementation of disparity or equity assessments profoundly difficult. As EO 13,985 itself notes, "Many Federal datasets are not disaggregated by race, ethnicity, gender, disability, income, veteran status, or other key demographic variables."

We studied agency action plans filed in response to the EO and conduct a comprehensive assessment across high-volume claims agencies that provide a wide range of government services (e.g., food assistance, farm subsidies, patents, tax refunds, loan guarantees), as well as all agencies filing equity action plans under the EO. Our analysis demonstrates how the privacy-bias tradeoff has undercut efforts to implement equity assessments. First, we show that the public response has been uniformly positive toward conducting equity assessments, with virtually no pushback on privacy grounds, with many agencies proposing substantial efforts. Second, we demonstrate that in practice, the challenges posed by the tradeoff have been profound. Twenty-one of 25 agencies note the lack of demographic information as a challenge. Agencies may be affirmatively prohibited from linking demographic records under privacy provisions. The Internal Revenue Service, for instance, has indicated that it would require statutory changes to be able to link to Census data to conduct an equity assessment [58]. Implementation of the Privacy Act's data minimization principle has led to pathological results: until 2022, for instance, the Food and Nutrition Service imputed race by "visual observation" by officials (e.g., the individual processing an application) when race was not self-reported to the Supplemental Nutrition Assitance Program [53]. Data minimization has meant that even where agencies want to acquire demographic information in principle, the legal, data infrastructure, and bureaucratic hurdles are severe. Third, we derive policy implications to address these barriers. Streamlining the approval for data collection for disparity assessments, restricting demographic data access to teams conducting an assessment (most ambitiously through the prototype National Secure Data Service (NSDS) [91] or the proposed National Artificial Intelligence Research Resource (NAIRR) [86]), and providing technical assistance to adapt the most appropriate methods would each enable disparity assessments, without seriously undermining individual privacy or the Privacy Act. Data minimization should not function as a license for blindness to disparities.

To our knowledge, this is the first paper to document the privacy-bias tradeoff at federal agencies. We make four distinct contributions. First, we conduct in-depth case studies to understand how federal agencies that affect large parts of the U.S. population have grappled with the privacy-bias tradeoff in practice. The federal experience with data minimization offers unique insights into the emerging tension between privacy and fairness. Second, we outline the range of distinct data approaches that have been taken — from record linkage, to amending forms, to commissioning a separate survey, to racial imputation methods — and discuss their legal and statistical challenges. Third, we identify the most common barriers to implementing equity assessments, which are centrally shaped by the data minimization approach taken under the Privacy Act of 1974, but also present a range of associated legal, data infrastructure, and bureaucratic land mines. Last, we provide a series of concrete and implementable policy recommendations to both protect privacy principles and enable equity – and algorithmic fairness – assessments.

We proceed as follows. Section 2 discusses primitive concepts and our research approach. Section 3 provides an assessment of data minimization at federal agencies, with detailed case studies in Appendix B. Section 4 discusses barriers emanating from restricted interpretations of privacy law, resistance by agencies and third parties, and fragmented data infrastructure. Section 5 concludes with implications.

2 DEFINITIONS, CONCEPTS, AND RESEARCH APPROACH

Definitions and Concepts. We begin from the premise that race and ethnicity are socially constructed [14, 78]. Some might take the position that the government, as a result, should never attempt to measure (or classify) race or ethnicity, but that position is refuted by EO 13,985, legal reporting requirements, as well as by the algorithmic fairness literature, which has focused on "fairness through awareness" [39]. Precisely because race and ethnicity have measurable disparate impacts on individuals, it is seen as critical to understand racial disparities in federal programs meant to equally benefit all. We also note that equity assessments are important across intersections of demographic characteristics (e.g., race and gender) [10, 28], but without some way to measure race or ethnicity, intersectional assessments remain impossible [54].

The federal approach to measuring race and ethnicity has varied over time and across agencies. The Social Security Administration (SSA), for instance, made changes to their "race/ethnicity codes" over decades [127]. The agency began collecting race data from enrollees in the 1930s [127]. Through 1980, enrollees self-identified as belonging to one of three categories: "White," "Negro," or "Other" [127]. Government-wide standards for collecting data on race and ethnicity were established by the OMB's 1977 "Standards for the Classification of Federal Data on Race and Ethnicity" and subsequently revised in 1997 [103, 104], but compliance with them has been uneven. On June 15, 2022, Dr. Karin Orvis, Chief Statistician of the United States, announced plans to review and revise the OMB standards to enable greater disaggregation of racial categories (e.g., representing subgroups of "Asian Americans"), but this effort is not expected to be completed until 2024 [113].

We use privacy throughout as shorthand to refer to informational privacy, namely the right of individuals to have a meaningful say in the way data about them is collected, stored, and used. There are multiple conceptions of privacy, such as the "right to be let

alone" [141], contextual integrity [92], and the fair information practice (FIP) principles-centered approach encoded in both the Privacy Act and E.U.'s General Data Protection Regulation. One of the key policy recommendations from the FIP perspective has been *data minimization*: that the collection of personal data is "adequate, relevant and limited to what is necessary in relation to the purposes for which they are processed" [45].

Research Approach. To explore the tension between bias assessment and privacy protections in general, we take a three-part approach. First, we conduct detailed case studies of demographic data collection, bias assessment, and privacy protections in large claims agencies. As there are over 300 federal agencies, we focus on high-volume claims agencies that affect large numbers of individuals. We consider all agencies with more than 1,000 claims opened or filed in 2013 [4] and more than $200 billion in 2023 budgetary resources [137]. We also include other high volume claims agencies, like the US Patent and Trademark Office (USPTO). This leaves us with a relatively comprehensive list of high-impact federal claims agencies. Second, for each agency, we select at least one high-impact program to evaluate in detail. For each program, we attempt to answer the following questions: What is the current practice of collecting race and ethnicity data? What is the current approach to estimating disparities or implementing EO 13,985? Have there been statutory or regulatory attempts to improve the ability of the agency to assess racial disparities? And if so, what barriers stand in the way of these improvements? All agencies, programs, and results are reported in Table 2. Third, we assess formal agency responses filed to the EO 13,985 and study stakeholder responses to the federal governments request for information for implementing EO 13,985.

3 THE STATE OF DATA MINIMIZATION AT FEDERAL AGENCIES

Responses to EO 13,985 from federal agencies and nongovernmental stakeholders suggest there is widespread support for conducting disparity assessments, but that existing demographic data poses serious challenges. Even before the EO was issued, numerous efforts proposed increasing data collection for disparity assessments. For example, Section 4302 of the Affordable Care Act (2010) requires national, federal data collection efforts to include race, ethnicity, sex, primary language, and disability status in order to "improve assessment of healthcare disparities" [59]. The Office of Civil Rights validates this "fairness through awareness" logic of bias assessment in healthcare, writing that "data collection is an important tool that can help covered entities to better serve their communities," specifically encouraging insurers to evaluate their services for different populations [101]. The Equal Credit Opportunity Act (ECOA), enacted in 1974, initially addressed discriminatory lending by severely restricting lenders' power to ask applicants for protected characteristics [2], but the Consumer Financial Protection Bureau (CFPB) eased this decades-long ban in 2017 to permit demographic data collection in more cases [19]. Table 1 summarizes case studies across federal agency programs.

3.1 Data Collection in 10 Large Claims Programs

Despite the positive reception to EO 13,985, Table 2 demonstrates a stark reality: demographic data is rarely, inconsistently, and poorly collected, and prior attempts to improve such collection ran into substantial barriers. Of the ten agencies studied, none systematically collects demographic data linked to program performance metrics and only two agencies have established a data linkage process that does not require direct data collection. For other agencies, the law or longstanding policy restricts data collection outright in some cases, while in others, the data collected is inadequate to support rigorous analysis of bias.

In four of the ten agencies studied, legal barriers prevent data collection for some or all programs. For instance, ECOA is an explicit legal barrier to demographic data collection across agencies; it precludes data collection for many agricultural loans facilitated by the USDA (See I11 in Table 2, where we use the letter to refer to column and the number to refer to the row as short hand) as well as small business loans like those issued by the pandemic-era Paycheck Protection Program (See Row 13 in Table 2). A 2008 Government Accountability Office (GAO) report concluded that ECOA's data limitations also complicate efforts to understand lending discrimination broadly and to identify specific lenders violating nondiscrimination requirements [145]. The Treasury Department's interpretation of privacy law and other relevant legislation mean the IRS only collects demographic attributes explicitly allowed in the tax code, excluding race, ethnicity, and sex (See I5 in Table 2) [58]. In a fifth case, the USPTO's standing practice also lacks demographic data questions in absence of a clear-cut statutory allowance for expanding data collection [61], despite an indirect call to "establish methods for studying the diversity of patent applicants" [68] (See Row 12 in Table 2).

Another three agencies attempt to collect demographic data, but lack the consistency and quality needed to support reliable bias assessments. The USDA is statutorily required by the 2008 Farm Bill to collect and publicly report demographic data for applicants to certain agricultural loans [107], but as one nonprofit, the Sustainable Agriculture and Food Systems Funders, notes, the website reporting USDA demographic data is "very badly out of date" and that "much of the data is missing" [133] (See I11 in Table 2). When applicants do not self-identify, USDA's demographic data sometimes relies on office employees' *visual assessment* of program applicants' race and ethnicity, despite long standing questions about the ethics and reliability of such an approach and a 2011 departmental regulation prohibiting such visual observation[2] [55]. Other USDA programs, such as the Supplemental Nutrition Assistance Program (SNAP), collect data more proactively but relied on visual observation when applicants did not self-report until 2021 [74] (See C10 in Table 2). Similarly, the Department of Veterans Affairs collects data from a variety of programs and sources but lacks "the complete and consistent collection of demographic data" that would support bias assessment [136] (See I3 in Table 2). In the case of Disaster Grants by

[2]A USDA report on the 2019 Market Facilitation Program notes that "departmental regulation prohibits the collection of race, ethnicity, and gender data based on a visual assessment, yet [Farm Services Agency] county office employees assigned race, ethnicity, and/or gender to producers through such means" [107].

the Department of Housing and Urban Development (HUD), only successful applicants provide demographic information, making an assessment of disparities in the application process impossible [57] (See I9 in Table 2).

The only agencies with a systematic approach to demographic data collection are SSA and HHS. Although neither directly collects race and ethnicity, SSA collaborates with other agencies, primarily Census, to link records from four different population surveys to determine race [79] (See C1 in Table 2). While HHS considers self-reported race and ethnicity to be the "gold standard" of accuracy [102], in practice it combines SSA's data with independent imputation to evaluate Medicare for racial bias (See C2 in Table 2). Table 1 enumerates the approaches to demographic data collection and comparative strengths and weaknesses. Agencies have six predominant approaches to collecting information about race and ethnicity: record linkage, voluntary direct data collection, mandatory direct data collection, imputation, survey-based random sampling, and visual observation. Subtle tradeoffs exist between these approaches. Survey-based random sampling, for instance, might untether bias assessment from legacy race and ethnicity categories, but stands to be significantly more costly than other approaches, particularly for small-sample demographic groups. Imputation is much more cost-effective, but conventional methods rely on restrictive assumptions [26] that are ideally validated on auxiliary datasets [41]. While these tradeoffs can be subtle, one approach is nearly universally deprecated, but still existent in federal datasets: visual observation [107].

3.2 Pre-Executive Order Attempts to Overcome Data Minimization

To understand how we arrived at this cacophony of race reporting, we now trace the barriers to improving data collection; improvements have been proposed for every program studied, except the Small Business Administration's Paycheck Protection Program (See Row 13 in Table 2), prior to the issuance of the EO. While these proposals vary in seriousness, approach, and progress towards implementation, the number of proposed expansions highlights the general consensus that demographic data collection is a worthwhile goal. In half of the case studies (See Rows 2-8 in Table 2), the agency itself is pursuing a serious data collection proposal. Some, like the HHS pilot program (See H2 in Table 2), are more limited in scope, while others, like Treasury's investment in imputation (See H5 in Table 2), intend to fully address the agency's data needs. Of the remaining case studies, HUD and USDA have initiated some agency efforts to collect demographic data. As SSA's bias assessment process is already relatively robust, changes to the long standing record linkage process with Census have only been floated in passing by small advocacy groups (See H1 in Table 2). In contrast, several bills have proposed demographic data collection for USPTO, but none has passed and no active effort to obtain data exists (See H12 in Table 2).

Across agencies, we observe five recurring classes of barriers, each influenced by privacy interests in its own way. First, legal restrictions directly prevent data collection, such as the Privacy Act or ECOA, or delay implementation, such as the Paperwork Reduction Act's notice and comment requirements (See I2 in Table 2). Second, fragmented or outdated technical infrastructure and a lack of technical expertise make systematic bias assessment challenging. Privacy measures to prevent unauthorized disclosure, while essential, further increase the technical resources required. Third, proposed data collectors, either the federal agency or a private third party, resist data collection, and in some instances we document evidence that this stems from the public relations, political, or litigation risk of uncovering bias in program administration. For instance, employers reporting to EEOC and lenders reporting to CFPB cite privacy and cost justifications, even in light of substantial protective measures proposed by the agency (see Appendix B.6 and B.3 for details). Fourth, federal agencies worry that asking respondents to provide demographic data will raise privacy concerns for respondents, and thus increase non-response rates. Finally, agencies lack the dedicated financial and personnel resources to implement improvements. Other barriers like technical limitations or complex legal review requirements contribute to the resource gap. (For convenience, we denote these enumerated barriers in Column I of 2 in parentheses.)

3.3 EO 13,985 Equity Action Plans

To gauge the extent to which federal agencies are actively grappling with the privacy-bias tradeoff in response to EO 13,985, we conducted a content analysis of all 25 available equity action plans filed in response to the order (see Appendix A for details). We assess (a) whether the availability of demographic data is recognized as a barrier, and (b) what concrete solutions (e.g., record linkage, form collection, visual observation, imputation, and sample surveys) are proposed to cure the data deficit. Table 3 synthesizes our findings.

The vast majority of agencies (21 of 25) highlight the lack of demographic data as a barrier to disparity assessments. The Treasury Department, for instance, states that the "right data to measure and advance equity is essential" but "challenges abound . . . since many federal datasets do not include race, ethnicity, or other key demographic variables." Similarly, the Department of Veterans' Affairs calls for expanded demographic data "to identify and eliminate disparities" [136]. The Federal Emergency Management Agency echoes this view: "The ability to collect demographic data . . . is imperative to achieving the intent and spirit of civil rights laws" [49]. While the wide acceptance of the role of demographic data in realizing anti-discrimination goals among federal agencies is certainly encouraging, far fewer equity plans supported their nominal recognition of the privacy-bias tradeoff with an concrete, actionable proposal of mechanisms to tackle it. Two agencies aim to run surveys; eight have committed to changing public-facing forms; four are proposing record linkage; and one is using imputation, but 13 agencies have at most partial or generic descriptions of any changes. Where an agency's plan noted–in varying detail–an interest in increasing data collection, most pointed to some variety of form collection. The NSF plans to display demographic questions upon entry into Research.gov and the Department of Education will begin requesting demographic information as part of FAFSA. The Treasury's limited demographic data sharing agreement with the Census Bureau, which supports cross tabulations of economic impact payments (EIP) data by race and ethnicity, is an affirmative

Table 1: Approaches to demographic data collection

Data collection method	Description	Strengths	Weaknesses	Self-reported	SORN	Notice & comment
Record linkage	Agency records are linked with existing race and ethnicity data (e.g., Census, Social Security Administration)	No new data collecting required. Responses already vetted through prior administrative records. Large population coverage.	Express statutory restrictions to protect privacy. Need for technical infrastructure and expertise. Nonresponse to administrative data. Record linkage error.	Y	Y	N
Form collection (voluntary)	Demographic data fields are included on registration forms for government services, but providing data is clearly voluntary and will not affect eligibility for benefits or services	Opt-out provides more assurance to respondents. Direct population of interest.	Nonresponse bias. Respondent time. Respondent concerns about use of race / ethnicity in program.	Y	N	Y
Form collection (mandatory)	Mandatory demographic data fields are included on registration forms for government services	Direct population of interest, with no nonresponse bias.	Lower response or participation rates. Respondent time. Respondent concerns about use of race / ethnicity in program.	Y	N	Y
Visual observation	Enumerators or program administrators assess what they believe to be the race, ethnicity, or gender of program participants.	Opt-out provides more assurance to respondents. Direct population of interest. Elimination of non-response bias without requiring respondents to answer.	Data reliability. Training of officials conducting observation. Available only for in-person enrollment.	N	N	N
Imputation	Demographic characteristics are inferred using statistical techniques, based on names, zip codes, and other predictive information	Can be performed for a near-full population. Only uses existing data and minimal public data.	Statistical bias. Results are more difficult to validate.	N	N	N
Survey-based random sampling	A random subset of program recipients (or of the population) is randomly sampled and surveyed by researchers to determine demographic characteristics	Reduces reporting burden. Reduces respondent concerns over misuse of demographic data. Flexibility as to reporting categories.	Expensive, as separate data collection required. Nonresponse bias. Need to oversample small demographic groups.	Y	N	Y

example of innovative inter-agency work to both actively recognise and concretely address the privacy-bias tradeoff.

Stakeholders, too, agree on the importance of collecting demographic data. A comprehensive review of each of the 531 responses to the OMB's Request for Information on EO 13,985 revealed nearly universal calls for increased data collection, sharing, and disaggregation of existing statistics. As Code for America summarized, "accurate and comprehensive demographic data" is essential because "you can't fix a problem you can't see" [75]. While a handful of organizations touch on privacy, cautioning that "data collection efforts must also be balanced with the importance of confidentiality and privacy, especially for vulnerable communities whose data may be disproportionately collected and shared," they still conclude that disaggregated data is "incredibly valuable, including as evidence of disparate impact, to help target resources, and to measure success" [87]. The Leadership Conference on Civil and Human Rights, the nation's oldest civil rights coalition, notes that even though "privacy is a real concern" that should be addressed, it "should not be used as a red herring to avoid collecting, disaggregating, or reporting data with the appropriate protections in place" [134]. This input highlights the potential for privacy concerns with government data collection to obstruct anti-discrimination efforts, while affirming that demographic data is essential and can be compatible with privacy interests.

4 STRUCTURAL BARRIERS TO EQUITY ASSESSMENTS

In this section, we synthesize the barriers we observed across case studies to discuss the statutory limits of federal privacy law and three types of obstacles associated with putting fairness assessments into practice: resistance from third parties, agencies' desire to maintain public trust, and infrastructural issues.

4.1 The Privacy Act

The Privacy Act of 1974 places significant limitations on the collection and use of personally identifiable information by government agencies. Passed in the wake of the Watergate scandal and amid growing concerns over government abuses of power and use of technology, the Act guards against the creation of a centralized federal database [109] through the adoption of a set of principles that were later enshrined into the Fair Information Practices: data minimization, purpose limitation, no disclosure without consent, rights of access and correction, and transparency (e.g., no secret data systems) [23]. Agencies can only collect information that is "relevant and necessary to accomplish a purpose of the agency" [77, §552(a)(e)(1)], and they are prohibited from disclosing personally identifiable information "to any person, or to another agency, except pursuant to a written request by, or with the prior written consent of, the individual to whom the record pertains" [77, §552(a)(a)(4)]. The statutory exceptions to the limits on disclosure often build on three general justifications: enabling statistical research [77, §552(a)(b)(5)], benefiting an agency's mandate (the agency has a "need to know" [77, §552(a)(b)(1)]), or "routine use" that is otherwise compatible with the purpose for which the data was collected [77, §552(a)(a)(7)]. Addressing bias is not explicitly acknowledged as a valid exception and is not easily justified through these standard avenues. More recently, the Confidential Information Protection and Statistical Efficiency Act of 2002 (CIPSEA) allowed "identifiable information" to be collected by federal agencies only for statistical purposes and under a pledge of confidentiality, strengthening the

Table 2: Data collection practices across federal agencies

Case Study		Current process		Proposed changes				I. Barriers
A. Sector	B. Agency	C. Collection method	D. Collection entity	E. Collection method	F. Collection entity	G. Source of proposal	H. Proposal status	
1. Social Security	SSA	Record linkage	Census	Form collection	SSA, hospitals	Suggestion from advocacy groups	No serious consideration or big backers [97]	Stopped collecting race data directly in 2002 after changes to the SSN assignment process, switched to Census data. Returning to form collection would be costly (5) [79].
2. Medicare	HHS	Record linkage, imputation	SSA (Census)	Form collection	SSA/CMS	HHS OIG recommendations	Pilot pending	The current approach relies on record linkage with SSA, supplemented by imputation to get race data [102]. HHS proposed a limited pilot to add race and ethnicity to Medicare Part C and D enrollment forms, but has been delayed by the notice and comment requirements of the PRA (1). Officials also point to a "feeling of risk" (3)[69].
3. Veterans benefits	VA	Form collection (voluntary)	Many (VA, Military, hospitals)	Form collection (mandatory)	VA	2021 bill from Sen. Hirono	Bill has not progressed	While the VA collects demographic data through a range of programs, it is not centralized, consistent, or complete (2). Their "Data for Equity" plan will synchronize data sets and address gaps [136]. VA officials find Sen. Hirono's bill to be misguided as demographic data is already collected, and resources are needed to integrate data sets and infrastructure (5) [16].
4. Veterans benefits	VA	Form collection (voluntary)	Many (VA, Military, hospitals)	Record linkage	VA	VA's 2022 Data for Equity plan	Recently launched, no reported programs yet	
5. Tax administration	Treasury	None	-	Imputation	Treasury	Office of Tax Analysis program	Implementation began in 2021 [5]	Can only collect data "necessary for tax administration" (1) and legal restrictions (1) on data sharing make it resource-intensive (5). Treasury officials worry mandatory form collection would reduce tax compliance (4) [58].
6. Wage discrimination enforcement	EEOC	None	-	Form collection (mandatory)	Employers	EEOC-initiated policy change	Will be implemented in July 2023	Hiring form EEO-1 collects demographic information from employers. Pay data was added to aid wage discrimination enforcement in 2016 but abruptly stayed in 2017. After lawsuits and a renewed research process, pay data collection will resume in July [60]. Some employers and elected representatives argue that pay data collection is a privacy risk and an undue burden (3) [84].
7. Mortgage lending	CFPB	Form collection [31]	Lenders	-	-	-	-	
8. Non-mortgage lending	CFPB	None [2]	-	Form collection	Lenders	CFPB rulemaking	Being written	ECOA bans data collection for non-mortgage lending (1). The CFPB grades banks on lending in low and middle income neighborhoods to proxy for racial redlining. The CFPB is writing regulations to use race and ethnicity directly to grade banks [143] and to require demographic data collection for small business loans [17]. Some lenders are pushing back (3) [8].
9. Community Development Block Grant Disaster Recovery	HUD	Form collection (partial)	HUD	Form collection (expanded)	HUD	2022 GAO recommendation [57]	HUD did not agree or disagree, but will research data improvements	HUD requires grantees to report race, ethnicity, and gender for those served, not for all applicants. Expanding data collection would require additional staffing, system infrastructure, and privacy protocols (2, 5) [57].
10. Food stamps	USDA	Visual observation, Form collection (voluntary) [53]	USDA employees	Form collection (voluntary), record linkage	State agencies	USDA suggestion	Not fleshed out	USDA is required to report beneficiaries' race and ethnicity to prevent discrimination, but lacks a robust alternative to visual observation, stopped in 2021, for missing data. They suggest using public education to raise response rates and state data or school records, where possible, to supplement self-reporting [53].
11. Farm subsidies	USDA	Visual observation, Form collection (voluntary) [15]	USDA employees	Form collection	USDA	2021 bill from Sen. Booker [15]	Bill has not progressed	Data collection is banned for some programs due to ECOA (1) and required for others. Collection efforts are disjointed, stymied by technical and procedural missteps (2, 5) [107]. Existing data reporting is inconsistent and incomplete [133].
12. Patents	USPTO	None [68], but one-off assessments via record linkage [135]	-	Form collection (voluntary)	USPTO	2021 bill from Sen. Schumer [126]	Amendment dropped when bill was passed [124]	Senators questioned the necessity of demographic data. Agency expressed privacy concerns with mandatory collection (4) and quality concerns with voluntary collection. [66]
13. Paycheck Protection Program	SBA	Form collection (voluntary)	Lenders	-	-	-	-	No modifications were considered in the PPP's three-year tenure.

Table 3: Review of agencies' equity action plans

	OPM	Interior	Treasury	SSA	GSA	DOJ	NSF	FEMA	HUD	Education	Labor	VA	Peace Corps	DHS	Transportation	Commerce	Energy	Defense	State	NASA	Agriculture	HHS	SBA	EPA	USAID
Stated need for demographic data?	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓		✓		✓	✓	✓	✓	✓	✓	✓	✓		
Sample surveys	■			■		■	■	■	■	■			■		■										
Form collection	■	■		■		■	■	■	■	■				■											
Record linkage	■	■	■	■		■					■	■													
Imputation		■	■																						
Visual observation																									

Note: Green indicates that a concrete plan was proposed and yellow that a partial plan or generic mention appears.

statistical research exception by allowing data sharing between statistical agencies (e.g., Census, Bureau of Economic Analysis, Bureau of Labor Statistics). We will focus on the restrictions on data sharing in this section.

The U.S. Commission on Evidence-Based Policymaking's 2018 Survey of Federal Agencies and Offices identified the Privacy Act as a major barrier to evidence-based policymaking. Of 79 respondents across various federal agencies, 47 cited "statutes prohibiting data sharing" as a barrier to data linkage [90, p. 16]. The only reason cited more frequently, by 66 respondents, was other "regulations and policies that make it difficult to link data" [90, p. 16]. Nineteen percent of respondents considered "legal limitations" to be the single most significant barrier they face in using data for evidence-building [90, p. 19]. The Privacy Act and other data protection provisions that provide additional guidance and obligations for compliance, such as the e-Government Act of 2002, place a substantial burden on data collection and sharing, and consequently, on efforts to identify and reduce bias in government programs. As maintaining public trust is essential for federal agencies to operate effectively, they take a conservative approach to data sharing when the law is unclear, contradictory, or silent on its legality [90, pp. 6-8]. Unfortunately, the complexity of privacy laws and exceptions create uncertainty about the precise restrictions on any given dataset.

In addition to the Privacy Act, which limits both disclosure of records without individual consent as well as record disclosures between agencies without written agreements, several other statutes[3] pose restrictions that must be reconciled to understand the legal barriers to combined datasets [90]. Agency practices can turn into "cognitive limits" [108, p. 140] functioning independently of the letter of the law [108, p. 140]. In addition, evaluating risk plays a substantial role in decisions to share data. Agencies are charged with managing risk to the organization but have been exhorted to extend their ambit to include risks to individuals as well [105, p.16].[4] This calculus shifts the weight against sharing data to minimize risk.

The IRS offers an instructive example: "[t]itles 13 [providing Census's privacy and confidentiality restrictions] and 26 [the Internal Revenue Code] of the United States Code limit the ability of Census and IRS to share data" and "Treasury officials report that laws protecting confidentiality prohibit IRS from acquiring demographic data from Census" [58, pp. 16-17] that could be used systematically to link data and evaluate bias. While Census and Treasury do enter project-based statistical research data sharing agreements, the necessary legal reviews require significant time and resources [58, p. 17]. One strength of imputation, Treasury's favored approach, is that it does not require large-scale data sharing, and thus legal review. The limitations imposed by privacy law and practice are summarized by a respondent to the Commission on Evidence-Based Policymaking's Survey of Federal Agencies and Offices, who says: "many agencies have restrictive requirements or restrictive interpretations of confidentiality laws and regulations that make it difficult to access valuable supplemental data." Another respondent goes further, asserting that "the most critical barrier to data exchange is legal and disclosure limitations" [90, p. 17]. Finally, as mentioned above, a legacy of the Privacy Act that continues to hamper data sharing efforts is the fear of centralized government data resources. While a decentralized approach may protect the public's privacy by making it difficult, if not impossible, for individuals to be tracked easily across agencies, this emphasis on decentralization reduced any perceived need for interoperable data infrastructure. Thus, inconsistencies in technical infrastructure proliferate and further obstruct data sharing, which in turn deprioritizes interoperability. The proposal for the National Secure Data Service (NSDS), for example, attempts to work around the concerns of a centralized database while enabling data sharing for evidence building:"[t]he Evidence Commission rejected a large-scale data warehouse model due to its untenable privacy risks and practical limitations for implementation. Instead, the experts encouraged the establishment of a National Secure Data Service as a shared service for conducting temporary data linkages for exclusively statistical purposes" [91].

4.2 External Resistance to Data Collection by Agencies

In cases where agencies consider expanding data collection requirements for third party service providers, these third parties have

[3] E.g., Title V of the e-Goverment Act, CIPSEA, the Family Educational Rights and Privacy Act (FERPA), etc.

[4] "When considering privacy risks, privacy programs shall consider the risks to an individual or individuals associated with the agency's creation, collection, use, processing, storage, maintenance, dissemination, disclosure, or disposal of their PII."

reacted with privacy concerns. For example, the CFPB issued policy guidance in 2018 regarding loan-level data collected under the Home Mortgage Disclosure Act (HMDA), which requires lenders to collect demographic data from mortgage applicants [20, p. 18]. The guidance recommended that loan-level data should be modified before public releases to prevent individual borrower re-identification. CFPB noted several industry comments which argued that privacy measures "did not sufficiently address" the risks of disclosure, but these comments "offered little evidence or analysis to support their views." Some industry commenters stated that the CFPB should only release aggregate-level data, or not release any data to protect borrowers' privacy. One commenter stated, "if there is 'any chance' that HMDA data could be used for criminal purposes, the benefits of disclosure could not outweigh the privacy risks." In contrast, consumer advocates argued that loan-level data has "long been publicly disclosed without any evidence the data has been used to harm applicants and borrowers." CFPB concluded that the risks—which are nonzero—"are justified by the benefits in light of HMDA's purposes" [20]. A similar tension when the Equal Employment Opportunity Commission (EEOC) considered collecting pay data in addition to demographic data to target wage discrimination. During the initial research process in 2012, employers' representatives expressed concerns about protecting individual privacy in aggregate data releases, so the EEOC re-examined statistical confidentiality standards to ensure tables with small cell-counts would be kept private [43]. After a multi-year, intensive research process, OMB decided in 2016 to begin collecting pay data. In 2017, OMB abruptly changed course, staying pay data collection on the grounds that it "lacks practical utility, is unnecessarily burdensome, and does not adequately address privacy and confidentiality issues" [116]. After worker groups sued, a federal judge reinstated pay data collection in 2019, finding OMB's decision arbitrary and capricious and reiterating the value of pay data for self-monitoring and enforcement purposes [129]. Even so, in 2020, Congresswoman Virginia Foxx (R-NC) said in a hearing "that the commission has no way of keeping [pay data] confidential" [84].

These cases illustrate the tension between protecting individual privacy and collecting and releasing data necessary for bias assessment. Even in cases where federal agencies institute technical privacy protections to mitigate concerns, the question of how privacy risks should be weighed against the benefits of data collection remains.

4.3 Structural Barriers to Direct Data Collection by Agencies

While data sharing with other agencies is cumbersome, direct data collection raises concerns that the public may respond negatively. Even if individual privacy is not substantively threatened and collection would be permitted by law, agencies worry the public may feel threatened due to unfamiliarity with existing privacy protections, a lack of awareness of the collective benefits of demographic data collection, or distrust of the data collectors. When the justification for data collection under privacy law is murky, agencies are especially likely to behave conservatively to maintain public trust. The USPTO sought public comment on a data collection proposal for bias assessment for a 2012 study on diversity among patent applicants. It found that "the ability of mandatory surveys to generate individual demographic diversity data of acceptable quality and reliability is in tension with the lack of public support for mandatory surveys due to privacy concerns under current law" [135, p. 3] One commenter noted that voluntary surveys would "reassure" respondents about their privacy [135, p.3].

Although voluntary survey questions may maintain respondents' sense of trust, they also pose the risk of lower response rates. A USDA report on the Market Facilitation Program (MFP), a program that distributed $25 billion to farmers hurt by retaliatory tariffs in 2018 and 2019, found that less than a third of recipients self-reported their race [99]. Minority groups often exhibit higher nonresponse rates for surveys in general [7, 50, 72] (though see Lee et al. 76), validating the USPTO's concern that voluntary response may produce "statistical bias arising from self-selection among respondents" if the non-response is not random [135, p.3]. CBAMS Survey and Focus Groups document high levels of mistrust in the government and public institutions, particularly among marginalized groups, offering one explanation for low response rates and suggesting disparate nonresponse bias [81]. Some have suggested offering monetary incentives to boost response rates for government surveys, though this solution comes with its own set of problems [128].

For these reasons, mandatory data collection does not appear to be a favored choice amongst agencies. As USDA moves away from using visual observation when respondents do not self-identify, it has suggested states should "encourage [participants] to self-identify and self-report" through education about the use of demographic data for bias assessment, and should find "other data sources or statistical tools to account for the times when participants choose not to self-identify" [53].

Mandatory data collection is seen to create risk with program participation. Tax experts, for instance, agree that adding demographic data questions to Form 1040 risks reduced tax compliance [58]. Even if tax auditors are denied access to race data, eliminating any possibility of express discrimination, interviewees concur that people may not file taxes if they perceive the government to be overreaching and potentially discriminating [58, p. 14]. Comments from a Treasury statistician suggest the need to manage public perception may affect Treasury particularly acutely; other agencies worry that even sharing data with Treasury could depress survey response rates by creating a perception that responses could be used to enforce tax compliance [115]. The effect of demographic questions on response rates to government surveys is difficult to predict and likely varies by context and demographic characteristic of interest. Based on research conducted by the Census Bureau, questions asking about sexual orientation and gender identity don't significantly depress responses [112], but serious concern about the safety, security, and integrity of the Census does exist, particularly amongst racial and ethnic minority groups [81]. Finally, a 2018 study by the Census Bureau identified the proposed reintroduction of a citizenship question on the 2020 census as a "major barrier" to participation, due to the political discourse surrounding immigration, and fears of retaliation against specific ethnic groups by the government [140]. Multiple studies suggest that a citizenship question would induce significant non-response [65]. Other agencies like the Center for Medicare Services intend to explore the

impact of demographic questions on non-response and program participation, but this research is not yet complete [102].

Direct data collection demonstrates the challenge of public trust and privacy. Generalized distrust of the federal government, ongoing concerns about government surveillance, and perhaps, lack of public awareness of federal privacy protections, especially in comparison to the private sector's lack of privacy regulation, all pose challenges to federal agencies navigating the requirements of the EO while maintaining public trust.

4.4 Fragmented Federal Infrastructure

A last major barrier to disparity assessments lies in the state of federal data infrastructure. Legacy systems, limited provisions for interoperability, lack of technical expertise, and administrative burdens can make data collection and sharing costly. Many scholars document the technical limitations that obstruct data sharing and linkage; O'Hara and Medalia specifically note a lack of staff, interoperability requirements, and over-specificity in funding "even when sharing is advantageous"[108].

One example of a deficit in technical infrastructure damaging data collection is the USDA's failed attempt to update demographic data collection for agricultural lending through the Market Facilitation Program. Even though a 2011 departmental regulation prohibited employees from using visual observation to determine race and ethnicity, the USDA's customer data management system continued to require that employees enter a value for demographic fields as late as 2019, so over two-thirds of race records for the program were still determined by employees' assessments. After the USDA realized the flaws in their data management system, they committed to remedying them, but still noted that fully updating their data management to make race and ethnicity optional would take months [107].

The Veterans Affairs administration faces similar obstacles; despite a desire to evaluate their programs using demographic data, the VA lacks the technical, personnel, and financial resources to update their fragmented infrastructure. VA data is maintained by a range of data stewards, leading to what some University of California San Francisco researchers call a "sometimes-confusing alphabet soup of data partners" [24]. In response to a bill proposing mandatory data collection [62], the VA noted that while demographic data is already collected, funds should be directed towards "improving existing collection, storage, management and analytics efforts" [16]. Rather than investing in a new form collection process, the VA's data officer calls for transforming the existing VA Profile system into a centralized data hub. The VA response to EO 13,985 recognizes that inconsistent data collection damages VA's "ability to assess where potential disparities or barriers exist" and calls for a "Data for Equity strategy…that will synchronize VA's data on health care, disability benefits, other veteran-facing services, and address data gaps in demographic information" [136]. Here, even with the desire to collect and use demographic data, a range of programs and data managers create a technical barrier.

The Commission on Evidence-Based Policymaking's survey of federal agency employees found that variation in agencies' technical infrastructure poses a significant challenge, noting specifically that agencies may struggle to "conduct disclosure reviews" and institute "disclosure avoidance protocols" [90, p.4] While statistical agencies are better equipped to collect, link, and analyze data while adhering to privacy standards, they are "less likely to view the purpose for which they collect data to be as a resource for evaluating programs" due to the restrictions of laws like CIPSEA, which limit disclosure of personally identifiable information for statistical agencies [90, p.4]. Beyond the well-documented technical and resource constraints on data management, bias assessment efforts can be specifically stymied by lacking technical infrastructure and expertise required to implement privacy protections.

5 SOLUTIONS

The problem at the heart of this paper is the privacy-bias tradeoff. Finding resolution inescapably requires a balance between how the U.S. government (a) protects individual privacy rights and (b) addresses structural problems of institutional bias. Ironically, the primary concerns motivating the passage of the Privacy Act—e.g., profiling and surveillance of individuals by the government—have proliferated in the private sector, providing a counterpoint of the perils of unchecked data collection and use [118]. We emphasize that our proposed solutions do not advocate for abandoning either data minimization or the Privacy Act. Congress amended the Act in 1988 with procedural safeguards to acknowledge the need of agencies to engage in some forms of record matching, while protecting privacy [118]. Similarly, our proposals attempt to enable the assessment of disparities in government programs while preserving the principles of the Privacy Act.

First, Congress should consider adding an exception to the Privacy Act that permits inter-agency record linkage specifically for bias assessment subject to the protections we suggest below. Alternatively, the Privacy Act's exceptions for "statistical research" and "routine uses" [95] can be interpreted to subsume bias assessment, but such an interpretation could restrict the use of such data for programmatic improvement. As Xiang [147] notes, the UK's Information Commissioner's Office issued guidance that demographic data should be collected for bias mitigation. The UK Data Protection Act supports an exception to the E.U.'s GDPR to allow data collection for bias assessment "with a view to enabling …equality to be promoted or maintained" [11]. While the collection of demographic data may not be strictly required to administer a program, it is necessary to ensure fair administration and to discharge legal obligations under EO 13,985. The Privacy Act's purpose specification requirement ([77, §552(a)(e)(3)(B-C]) would obligate an agency to disclose on a form the "principal purpose or purposes for which the information is intended to be used," as well routine uses, which would supplement our fifth recommendation below. Inter-agency record sharing for bias assessment has the fewest methodological challenges of available approaches (see Table 1) and would solve the first order barrier for agencies to conduct disparity assessments.

Second, while such inter-agency data sharing would enable disparity assessments, we also recommend institutional protections to guard against misuse of demographic data. The internal "separation of functions" (e.g., between investigatory and adjudicatory functions) has long been a mainstay of administrative law [12, 36] and was one of the original recommendations from the 1973 Department of Health, Education, and Welfare report that preceded

the Act[110]. A similar internal separation of functions should be available only to offices conducting the disparity assessment, and such offices should be distinct from offices processing claims [83, 132]. This separation would be consistent with the Privacy Protection Study Commission's conclusion that "no record . . . collected or maintained for a research or statistical purpose . . . may be used in individually identifiable form to make any decision or take any action directly affecting the individual to whom the record pertains" [100, 111]. For instance, Census demographic information could be made exclusively available to the Treasury Department's Office of Tax Analysis, not IRS's audit units, which is consistent with the current approach by Treasury [58]. Such separation of functions would insulate sensitive data from enforcement offices, building public trust and ensuring fair and equal treatment.

Third, to overcome substantial resource and data infrastructure challenges, Congress should increase support of initiatives like the NSDS [91], the NAIRR [63, 86], and other mechanisms to enable privacy-protecting sharing of administrative data for disparity assessments [90, 108]. The CHIPS and Science Act of 2022 funded a limited five-year trial of NSDS demonstration projects and the National Science Foundation has recently attempted to broker partnerships between academic researchers and agencies to implement EO 13,985 [88]. Such initiatives have significant potential to address data infrastructure, computing, and human capital gaps to conduct disparity assessments. Our case studies also revealed numerous instances where the apparent agency resistance stemmed from lack of technical resources to incorporate demographic data into agency systems. Congress and the President should explore mechanisms, such as GSA's 18F consultancy, the U.S. Digital Corps, or the U.S. Digital Service, to provide technical assistance for IT modernization to incorporate demographic data collection and restrict access to teams requiring that information.

Fourth, Congress should amend the Paperwork Reduction Act to provide a streamlined process for capturing demographic data on federal forms or running auxiliary surveys. Particularly as Census and OMB update their data standards for race reporting, there is no need for each agency to undergo a separate OMB review process and notice and comment to collect race and ethnicity information in a manner already approved as a general data standard. To be sure, public input can be valuable, but procedural requirements can significantly impede the ability of agencies to collect information relevant to assessing racial disparities.

Fifth, in developing data strategies for disparity assessments, federal agencies should expressly address public concerns about forms, such as concerns about invasiveness and reduction of voluntary compliance [58, p.14-15]. Agencies should strategically select from the options enumerated in Table 1 to develop the most appropriate data strategy and communicate their rationale. For example, the USDA Food and Nutrition Service noted that programs "should continue explaining the importance of this data to participants as they encourage them to self-identify and self-report" [74].[5] Forms should clearly and simply explain how demographic data will be used and protected. For instance, if IRS were to collect race and ethnicity, it should clarify that such information would not be used to select audits.

Because our research reveals substantial institutional challenges, our recommendations focus on those dimensions. Privacy enhancing technologies such as differential privacy and secure multi-party computation are of course quite important here for enabling secure and private inter-agency data sharing, but we emphasize that they are unlikely a complete solution. Other agencies may follow the Census Bureau's adoption of differential privacy in 2020 [18], but Drechsler [38] notes that public sector requirements (e.g., for reproducibility, final data users, and data sharing) can be "fundamentally different from [those] in industry." Government data, for instance, is intended to be used over many decades, making the calculation of privacy budgets on a query system challenging. Privacy enhancing technologies may and should, of course, still be adapted consistent with our proposals that are aimed to enhance the current data deficit in demographic data for disparity assessments.

6 CONCLUSION

Nearly fifty years after the passage of the Privacy Act, the law remains both an exemplar of government privacy protections as well as a third rail in evolving privacy discussions. Over the years, the Act has been strengthened, and there are no serious discussions to weaken it. Ongoing concerns regarding the government's efforts to purchase data about individuals from the private sector has led to calls to eliminate loopholes [121]. Data minimization as policy is broadly viewed as a success; proposed rulemaking by the Federal Trade Commission in 2022 [29] as well as proposed bipartisan privacy legislation both embraced its adoption by the private sector [119]. And yet, our analysis highlights weaknesses with this half-century experiment. Concerns about exposure and knowledge of individuals by the government has resulted in a lack of visibility into the impact of policymaking on subgroups. We achieved individual privacy at the expense of collective knowledge, yielding conditions ripe for allowing disparate impacts to proliferate unchecked.

ACKNOWLEDGMENTS

We thank the Public Interest Technology University Network (PIT-UN) for generous support and Dawson Verley for research assistance. We are grateful to Chris Hoofnagle for his helpful comments.

REFERENCES

[1] [n. d.]. Questions and Answers Related to CACFP 11-2021, SFSP 07-2021 Collection of Race and Ethnicity Data by Visual Observation and Identification in the Child and Adult Care Food Program and Summer Food Service Program – Policy Rescission.
[2] 1974. Equal Credit Opportunity Act. https://www.ecfr.gov/current/title-12/chapter-X/part-1002/section-1002.5
[3] 2003. Nondiscrimination in Federally-Assisted Programs of the Department of Agriculture - Effectuation of Title VI of the Civil Rights Act of 1964. https://www.ecfr.gov/current/title-7/subtitle-A/part-15#p-15.5(b)
[4] ACUS and Stanford Law School. [n. d.]. Caseload Statistics. http://acus.law.stanford.edu/reports/caseload-statistics
[5] Wally Adeyemo and Lily Batchelder. 2021. Advancing Equity Analysis in Tax Policy. https://home.treasury.gov/news/featured-stories/advancing-equity-analysis-in-tax-policy
[6] Sushant Agarwal. 2021. Trade-Offs between Fairness and Privacy in Machine Learning. In *IJCAI 2021 Workshop on AI for Social Good*. https://crcs.seas.harvard.edu/publications/trade-offs-between-fairness-and-privacy-machine-learning.

[5] Similarly, in response to recommendations from the HHS Office of Inspector General, the Center for Medicare and Medicaid Services (CMS) wrote, "It is important that enrollees understand the value of [demographic] data and how the data will be utilized." [102, 20]

[7] Nanna Ahlmark, Maria Holst Algren, Teresa Holmberg, Marie Louise Norredam, Signe Smith Nielsen, Astrid Benedikte Blom, Anne Bo, and Knud Juel. 2015. Survey nonresponse among ethnic minorities in a national health survey – a mixed-method study of participation, barriers, and potentials. *Ethnicity & Health* 20, 6 (Nov. 2015), 611–632. https://doi.org/10.1080/13557858.2014.979768 Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/13557858.2014.979768.

[8] American Bankers Association, Bank Policy Institute, Consumer Bankers Association, Housing Policy Council, and Mortgage Bankers Association. 2019. Docket No. CFPB-2019-0020; Home Mortgage Disclosure (Regulation C) Data Points and Coverage. https://www.aba.com/-/media/documents/comment-letter/joint-ltr-hmda-residential-101519.pdf?rev=1546032f15b14cd19d61549d06334ed2

[9] McKane Andrus, Elena Spitzer, Jeffrey Brown, and Alice Xiang. 2021. What We Can't Measure, We Can't Understand: Challenges to Demographic Data Procurement in the Pursuit of Fairness. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT '21)*. Association for Computing Machinery, New York, NY, USA, 249–260. https://doi.org/10.1145/3442188.3445888

[10] McKane Andrus and Sarah Villeneuve. 2022. Demographic-reliant algorithmic fairness: characterizing the risks of demographic data collection in the pursuit of fairness. In *2022 ACM Conference on Fairness, Accountability, and Transparency*. 1709–1721.

[11] The National Archives. 2018. Data Protection Act 2018. https://www.legislation.gov.uk/ukpga/2018/12 Publisher: Statute Law Database.

[12] Michael Asimow. 1981. When the Curtain Falls: Separation of Functions in the Federal Administrative Agencies. *Columbia Law Review* 81, 4 (1981), 759–820.

[13] Eugene Bagdasaryan and Vitaly Shmatikov. 2019. Differential Privacy Has Disparate Impact on Model Accuracy. *arXiv:1905.12101 [cs, stat]* (Oct. 2019). http://arxiv.org/abs/1905.12101 arXiv: 1905.12101.

[14] Ruha Benjamin. 2019. Race after technology: Abolitionist tools for the new jim code. *Social forces* (2019).

[15] Cory Booker. 2021. Farm Subsidy Transparency Act of 2021. https://www.congress.gov/bill/117th-congress/senate-bill/1980/text

[16] Aaron Boyd. 2021. VA Officials and Lawmakers Have (Different) Issues With Push to Collect More Veteran Data. *Nextgov.com* (Oct. 2021). https://www.nextgov.com/analytics-data/2021/10/va-officials-and-lawmakers-have-different-issues-push-collect-more-veteran-data/185936/

[17] Consumer Financial Protection Bureau. 2021. Small Business Lending Data Collection under the Equal Credit Opportunity Act (Regulation B). https://www.consumerfinance.gov/rules-policy/rules-under-development/small-business-lending-data-collection-under-equal-credit-opportunity-act-regulation-b/

[18] US Census Bureau. 2020. Understanding Differential Privacy. https://www.census.gov/programs-surveys/decennial-census/decade/2020/planning-management/process/disclosure-avoidance/differential-privacy.html Section: Government.

[19] Bureau of Consumer Financial Protection. 2017. *Amendments to Equal Credit Opportunity Act (Regulation B) Ethnicity and Race Information Collection.* Technical Report 82 FR 45680. The Bureau of Consumer Financial Protection. 65 pages. https://files.consumerfinance.gov/f/documents/201709_cfpb_final-rule_regulation-b.pdf

[20] Bureau of Consumer Financial Protection. 2018. *Disclosure of Loan-Level HMDA Data.* Final policy guidance. CFPB-2017-0025. Bureau of Consumer Financial Protection. https://s3.amazonaws.com/files.consumerfinance.gov/f/documents/HMDA_Disclosure_FPG_--_Final_12.21.2018_for_website_with_date.pdf

[21] Bureau of Consumer Financial Protection. 2023. *Small Business Lending Data Collection under the Equal Credit Opportunity Act (Regulation B).* Technical Report RIN 3170-AA09. Bureau of Consumer Financial Protection. https://files.consumerfinance.gov/f/documents/cfpb_1071-final-rule.pdf

[22] Sylvia Burwell. 2014. *Guidance for Providing and Using Administrative Data for Statistical Purposes.* Memorandum for the Heads of Executive Departments and Agencies M-14-06. Office of Management and Budget. https://obamawhitehouse.archives.gov/sites/default/files/omb/memoranda/2014/m-14-06.pdf#page=16

[23] Electronic Privacy Information Center. [n. d.]. The Privacy Act of 1974. https://epic.org/the-privacy-act-of-1974/

[24] UCSF Pepper Center. 2023. Department of Veterans Affairs (VA) Data. https://peppercenter.ucsf.edu/department-veterans-affairs-va-data

[25] Hongyan Chang and Reza Shokri. 2021. On the Privacy Risks of Algorithmic Fairness. In *2021 IEEE European Symposium on Security and Privacy (EuroS&P)*. 292–303. https://doi.org/10.1109/EuroSP51992.2021.00028

[26] Jiahao Chen, Nathan Kallus, Xiaojie Mao, Geoffry Svacha, and Madeleine Udell. 2019. Fairness under unawareness: Assessing disparity when protected class is unobserved. In *Proceedings of the conference on fairness, accountability, and transparency*. 339–348.

[27] Miranda Christ, Sarah Radway, and Steven M. Bellovin. 2022. Differential Privacy and Swapping: Examining De-Identification's Impact on Minority Representation and Privacy Preservation in the U.S. Census. In *2022 IEEE Symposium on Security and Privacy (SP)*. IEEE, San Francisco, CA, USA, 457–472. https://doi.org/10.1109/SP46214.2022.9833668

[28] P.H. Collins and S. Bilge. 2020. *Intersectionality.* Polity Press. https://books.google.com/books?id=fyrfDwAAQBAJ

[29] Federal Trade Commission. 2022. Commercial Surveillance and Data Security Rulemaking. https://www.ftc.gov/legal-library/browse/federal-register-notices/commercial-surveillance-data-security-rulemaking

[30] Consumer Financial Protection Bureau. 2014. *Home Mortgage Disclosure (Regulation C).* Proposed Rule 12 CFR Part 1003. Consumer Financial Protection Bureau. https://www.regulations.gov/document/CFPB-2014-0019-0001

[31] Consumer Financial Protection Bureau. 2021. Home Mortgage Disclosure Act (HMDA) Data. https://www.consumerfinance.gov/data-research/hmda/

[32] Sam Corbett-Davies, Emma Pierson, Avi Feller, Sharad Goel, and Aziz Huq. 2017. Algorithmic Decision Making and the Cost of Fairness. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, Halifax NS Canada, 797–806. https://doi.org/10.1145/3097983.3098095

[33] Tadlock Cowan and Jody Feder. 2013. The Pigford Cases: USDA Settlement of Discrimination Suits by Black Farmers. (May 2013), 14.

[34] Julie-Anne Cronin, Portia DeFilippes, and Robin Fisher. 2023. Tax Expenditures by Race and Hispanic Ethnicity: An Application of the U.S. Treasury Department's Race and Hispanic Ethnicity Imputation. https://home.treasury.gov/system/files/131/WP-122.pdf

[35] Pete Daniel. 2013. *Dispossession: Discrimination Against African American Farmers in the Age of Civil Rights* (1 ed.). The University of North Carolina Press.

[36] Kenneth Culp Davis. 1948. Separation of Functions in Administrative Agencies. *Harvard Law Review* 61, 3 (1948), 389–418. http://www.jstor.org/stable/1335525

[37] Day One Project. 2021. *Transition Document for the United States Patent and Trademark Office.* Technical Report. Day One Project. https://www.dayoneproject.org/ideas/transition-document-for-the-united-states-patent-and-trademark-office/

[38] Jörg Drechsler. 2023. Differential Privacy for Government Agencies—Are We There Yet? *J. Amer. Statist. Assoc.* 0, ja (Jan. 2023), 1–24. https://doi.org/10.1080/01621459.2022.2161385 Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/01621459.2022.2161385.

[39] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. 2012. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*. 214–226.

[40] Cynthia Dwork and Jing Lei. 2009. Differential privacy and robust statistics. In *Proceedings of the forty-first annual ACM symposium on Theory of computing*. 371–380.

[41] Hadi Elzayn, Evelyn Smith, Tom Hertz, Arun Ramesh, Robin Fisher, Daniel E Ho, and Jacob Goldin. 2023. Measuring and Mitigating Racial Disparities in Tax Audits. https://dho.stanford.edu/wp-content/uploads/IRS_Disparities.pdf

[42] Equal Employment Opportunity Commission. [n. d.]. What You Should Know: The National Academies' Evaluation of Compensation Data Collected Through the EEO-1 Form. https://www.eeoc.gov/wysk/what-you-should-know-national-academies-evaluation-compensation-data-collected-through-eeo-1

[43] Equal Employment Opportunity Commission. 2016. *Agency Information Collection Activities: Revision of the Employer Information Report (EEO-1) and Comment Request.* 81 FR 5113. Equal Employment Opportunity Commission. https://www.federalregister.gov/documents/2016/02/01/2016-01544/agency-information-collection-activities-revision-of-the-employer-information-report-eeo-1-and

[44] Equal Employment Opportunity Commission. 2019. *Agency Information Collection Activities: Existing Collection.* Notice of Information Collection 84 FR 48138. Equal Employment Opportunity Commission. https://www.federalregister.gov/documents/2019/09/12/2019-19767/agency-information-collection-activities-existing-collection

[45] European Parliament. 2016. Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation). http://data.europa.eu/eli/reg/2016/679/oj/eng Legislative Body: EP, CONSIL.

[46] Executive Office of the President. 2020. *Promoting the Use of Trustworthy Artificial Intelligence in the Federal Government.* Technical Report E.O. 13960. Executive Office of the President. https://www.federalregister.gov/documents/2020/12/08/2020-27065/promoting-the-use-of-trustworthy-artificial-intelligence-in-the-federal-government

[47] Executive Office of the President. 2023. *Further Advancing Racial Equity and Support for Underserved Communities Through the Federal Government.* Technical Report E.O. 14,091. https://www.federalregister.gov/documents/2023/02/22/2023-03779/further-advancing-racial-equity-and-support-for-underserved-communities-through-the-federal

[48] Executive Order On Advancing Racial Equity and Support for Underserved Communities Through the Federal Government 2021. Executive Order On Advancing Racial Equity and Support for Underserved Communities Through the Federal Government. https://www.whitehouse.gov/briefing-room/presidential-actions/2021/01/20/executive-order-advancing-racial-equity-and-support-for-underserved-communities-through-the-federal-government/

[49] Federal Emergency Management Agency. 2022. *Federal Emergency Management Agency Equity Action Plan.* Technical Report. Federal Emergency Management Agency. https://www.fema.gov/sites/default/files/documents/fema_equity-action-plan.pdf

[50] Remco Feskens, Joop Hox, Gerty Lensvelt-Mulders, and Hans Schmeets. 2006. Collecting Data among Ethnic Minorities in an International Perspective. *Field Methods* 18, 3 (Aug. 2006), 284–304. https://doi.org/10.1177/1525822X06288756 Publisher: SAGE Publications Inc.

[51] Ferdinando Fioretto, Cuong Tran, Pascal Van Hentenryck, and Keyu Zhu. 2022. Differential Privacy and Fairness in Decisions and Learning Tasks: A Survey. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence*. 5470–5477. https://doi.org/10.24963/ijcai.2022/766 arXiv:2202.08187 [cs].

[52] Benjamin Fish, Jeremy Kun, and ?dám D. Lelkes. 2016. A Confidence-Based Approach for Balancing Fairness and Accuracy. In *Proceedings of the 2016 SIAM International Conference on Data Mining (SDM)*. Society for Industrial and Applied Mathematics, 144–152. https://doi.org/10.1137/1.9781611974348.17

[53] Food and Nutrition Service. 2022. Proposed Rule: SNAP-Revision of Civil Rights Data Collection Methods. https://www.fns.usda.gov/snap/fr-062722

[54] Robert Gellman. 2014. Fair Information Practices: A Basic History. *SSRN Electronic Journal* (2014). https://doi.org/10.2139/ssrn.2415020

[55] Government Accountability Office. 2008. *Recommendations and Options to Address Management Deficiencies in the Office of the Assistant Secretary for Civil Rights.* Report to Congressional Requesters GAO-09-62. Government Accountability Office. https://www.gao.gov/assets/gao-09-62.pdf

[56] Government Accountability Office. 2019. *Information on Credit and Outreach to Socially Disadvantaged Farmers and Ranchers Is Limited.* Technical Report GAO-19-539. Government Accountability Office. https://www.gao.gov/assets/gao-19-539.pdf

[57] Government Accountability Office. 2022. *Better Data Are Needed to Ensure Equitable Delivery of HUD Block Grant Funds to Vulnerable Populations.* Technical Report GAO-22-105548. Government Accountability Office. https://www.gao.gov/assets/gao-22-105548-highlights.pdf

[58] Government Accountability Office. 2022. *Lack of Data Limits Ability to Analyze Effects of Tax Policies on Households by Demographic Characteristics.* Report to Congressional Committees GAO-22-104553. Government Accountability Office. 101 pages. https://www.gao.gov/assets/gao-22-104553.pdf

[59] Health and Human Services. [n. d.]. *Improving Data Collection to Reduce Health Disparities.* Technical Report. https://minorityhealth.hhs.gov/assets/pdf/checked/1/Fact_Sheet_Section_4302.pdf

[60] Akerman LLP-Tiffany D. Hendricks. 2023. 2022 EEO-1 Component 1 Data Collection Now Set to Begin Mid-July 2023. https://www.lexology.com/library/detail.aspx?g=38921489-f809-47e3-9ff9-2581c0954f09

[61] Mazie Hirono. 2021. Inventor Diversity for Economic Advancement Act of 2021. https://www.congress.gov/bill/117th-congress/senate-bill/632

[62] Mazie K. Hirono. 2021. Every Veteran Counts Act of 2021. http://www.congress.gov/ Archive Location: 2021/2022.

[63] Daniel Ho, Jennifer King, Russell Wald, and Christopher Wan. 2021. *Building a National AI Research Resource: A Blueprint for the National Research Cloud.* Technical Report. Human-Centered Artificial Intelligence. https://hai.stanford.edu/sites/default/files/2022-01/HAI_NRCR_v17.pdf

[64] Kosuke Imai, Santiago Olivella, and Evan T. R. Rosenman. 2022. Addressing census data problems in race imputation via fully Bayesian Improved Surname Geocoding and name supplements. *Science Advances* 8, 49 (Dec. 2022). https://doi.org/10.1126/sciadv.adc9824

[65] Matthew D Ingber, Andrew J Pincus, Michael B Kimberly, and Colleen M Campbell. [n. d.]. United States Department of Commerce vs. State of New York. https://www.supremecourt.gov/DocketPDF/18/18-966/95014/20190401170915326_18-966.bsac.pdf

[66] IPWatchdog. 2021. IDEA Act Passed Out of Senate Judiciary Committee. *IPWatchdog.com* (April 2021). https://ipwatchdog.com/2021/04/29/idea-act-passed-senate-judiciary-committee/id=132917/

[67] Matthew Jagielski, Michael Kearns, Jieming Mao, Alina Oprea, Aaron Roth, Saeed Sharifi Malvajerdi, and Jonathan Ullman. 2019. Differentially Private Fair Learning. In *Proceedings of the 36th International Conference on Machine Learning*. PMLR, 3000–3008. https://proceedings.mlr.press/v97/jagielski19a.html ISSN: 2640-3498.

[68] Shontavia Johnson. 2019. The Colorblind Patent System and Black Inventors. *Landslide* 11, 4 (April 2019). https://www.americanbar.org/groups/intellectual_property_law/publications/landslide/2018-19/march-april/colorblind-patent-system-black-inventors/#72

[69] John Hewitt Jones. 2023. HHS chief data officer: Federal employees face outsized barriers to data sharing. *FedScoop* (Jan. 2023). https://www.fedscoop.com/hhs-chief-data-officer-federal-employees-face-outsized-barriers-to-data-sharing/

[70] Faisal Kamiran and Toon Calders. 2012. Data preprocessing techniques for classification without discrimination. *Knowledge and Information Systems* 33, 1 (Oct. 2012), 1–33. https://doi.org/10.1007/s10115-011-0463-8

[71] Ron D Katznelson. 2021. The IDEA Act is a Bad Idea. *IPWatchdog.com* (March 2021). https://ipwatchdog.com/2021/03/24/idea-act-bad-idea/id=131357/

[72] Hyun-Jun Kim and Karen I. Fredriksen-Goldsen. 2013. Nonresponse to a Question on Self-Identified Sexual Orientation in a Public Health Survey and Its Relationship to Race and Ethnicity. *American Journal of Public Health* 103, 1 (Jan. 2013), 67–69. https://doi.org/10.2105/AJPH.2012.300835 Publisher: American Public Health Association.

[73] Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. 2017. Inherent Trade-Offs in the Fair Determination of Risk Scores. (2017), 23 pages. https://doi.org/10.4230/LIPICS.ITCS.2017.43 Artwork Size: 23 pages Medium: application/pdf Publisher: Schloss Dagstuhl - Leibniz-Zentrum fuer Informatik GmbH, Wadern/Saarbruecken, Germany.

[74] Angela Kline and Roberto Contreras. 2021. *Collection of Race and Ethnicity Data by Visual Observation and Identification in the Child and Adult Care Food Program and Summer Food Service Program – Policy Rescission.* Technical Report CACFP 11-2021, SFSP 07-2021. U.S. Department of Agriculture. https://www.fns.usda.gov/cn/Race-and-Ethnicity-Data-Policy-Rescission

[75] Ryan Ko. 2021. *Comment from Code for America.* Technical Report OMB-2021-0005-0428. 7 pages. https://www.regulations.gov/comment/OMB-2021-0005-0428

[76] Sunghee Lee, Karen I. Fredriksen-Goldsen, Colleen McClain, Hyun-Jun Kim, and Z. Tuba Suzer-Gurtekin. 2018. Are Sexual Minorities Less Likely to Participate in Surveys? An Examination of Proxy Nonresponse Measures and Associated Biases with Sexual Orientation in a Population-based Health Survey. *Field Methods* 30, 3 (Aug. 2018), 208–224. https://doi.org/10.1177/1525822X18777736 Publisher: SAGE Publications Inc.

[77] Legislative History of Privacy Act 1976. Legislative History of the Privacy Act of 1974. https://lccn.loc.gov/2011525305

[78] Ian F Haney Lopez. 1994. The social construction of race: Some observations on illusion, fabrication, and choice. *Harv CR-CLL Rev.* 29 (1994), 1.

[79] Patricia Martin. 2016. Why Researchers Now Rely on Surveys for Race Data on OASDI and SSI Programs: A Comparison of Four Major Surveys. https://www.ssa.gov/policy/docs/rsnotes/rsn2016-01.html

[80] Massachusetts Law Reform Institute. 2021. *Comment from Massachusetts Law Reform Institute.* Technical Report USDA-2021-0006-0261. https://www.regulations.gov/comment/USDA-2021-0006-0261

[81] Kyley McGeeney. 2020. *2020 Census Barriers, Attitudes, and Motivators Study Survey Report.* Technical Report. 381 pages.

[82] Jena McGregor. 2019. An Obama-era rule to collect worker pay data is headed for the chopping block. *The Washington Post* (Sept. 2019). https://www.washingtonpost.com/business/2019/09/13/an-obama-era-rule-collect-worker-pay-data-is-headed-chopping-block/

[83] Geoffrey P Miller. 1986. Independent agencies. *The Supreme Court Review* 1986 (1986), 41–97.

[84] J. Edward Moreno. 2022. EEOC Chair Makes Case on Hill for Access to Employer Pay Data. *Bloomberg Law* (April 2022). https://news.bloomberglaw.com/daily-labor-report/eeoc-chair-makes-case-on-hill-for-access-to-employer-pay-data

[85] Engineering National Academies of Sciences and Medicine. 2022. *Evaluation of Compensation Data Collected Through the EEO-1 Form.* The National Academies Press, Washington, DC. https://doi.org/10.17226/26581

[86] National Artificial Intelligence Research Resource Task Force. 2023. *Strengthening and Democratizing the U.S. Artificial Intelligence Innovation Ecosystem: An Implementation Plan for a National Artificial Intelligence Research Resource.* Technical Report. National Artificial Intelligence Research Resource Task Force. https://www.ai.gov/wp-content/uploads/2023/01/NAIRR-TF-Final-Report-2023.pdf

[87] National Legal Aid & Defender Association. 2021. *Comment from National Legal Aid & Defender Association.* Technical Report OMB-2021-0005-0520. National Legal Aid & Defender Association. https://www.regulations.gov/comment/OMB-2021-0005-0520

[88] National Science Foundation. 2022. *The Analytics for Equity Initiative.* Technical Report. https://nsf-gov-resources.nsf.gov/2022-06/Summary20of%20the%20Analytics%20for%20Equity%20Initiative.pdf

[89] Craig Nazzaro, Brad Rustin, and Elizabeth DeVos. 2017. Privacy Concerns Remain as HMDA Implementation Date Arrives. *https://www.bankdirector.com/* (Dec. 2017). https://www.bankdirector.com/index.php/issues/risk/privacy-concerns-remain-hmda-implementation-date-arrives/ Section: Issues.

[90] Nick Hart and Kody Carmody. 2018. *Barriers to Using Government Data: Extended Analysis of the U.S. Commission on Evidence-Based Policymaking's Survey of Federal Agencies and Offices.* Technical Report. Bipartisan Policy Center, Washington, D.C.

[91] Nick Hart and Nancy Potok. 2022. *A Blueprint for Implementing the National Secure Data Service: Initial Governance and Administrative Priorities for the National Science Foundation.* Technical Report. Data Foundation. 28 pages.

[92] Helen Fay Nissenbaum. 2010. *Privacy in context: technology, policy, and the integrity of social life.* Stanford Law Books, Stanford, Calif.

[93] Office of Congresswoman Nydia Velazquez. 2018. Lawmakers Ask USPTO to Take Steps to Diversify Patent Holders. https://velazquez.house.gov/media-

center/press-releases/lawmakers-ask-uspto-take-steps-diversify-patent-holders

[94] Office of Cory Booker. 2021. Booker, Rush Introduce Landmark Bicameral Bill To Lift Veil Of Secrecy And Ensure Transparency In Farm Subsidy Recipients. https://www.booker.senate.gov/news/press/booker-rush-introduce-landmark-bicameral-bill-to-lift-veil-of-secrecy-and-ensure-transparency-in-farm-subsidy-recipients

[95] U.S. Department of Justice. 2020. Overview of The Privacy Act of 1974. https://www.justice.gov/opcl/overview-privacy-act-1974-2020-edition

[96] Office of Mazie K. Hirono. 2021. Hirono, Tillis, Velázquez, Stivers Introduce Bipartisan, Bicameral Bill to Close the Patent Gap Faced by Women, Minorities. https://www.hirono.senate.gov/news/press-releases/hirono-tillis-velzquez-stivers-introduce-bipartisan-bicameral-bill-to-close-the-patent-gap-faced-by-women-minorities

[97] Community Legal Services of Philadelphia. 2020. Collection and Regular Reports on Race and Ethnicity Data within Social Security Programs. https://clsphila.org/wp-content/uploads/2022/06/Letter-to-SSA-Re-Race-and-Ethnicity-Data.pdf

[98] White House Office of Science and Technology Policy. 2022. *Blueprint for an AI Bill of Rights.* Whitepaper. White House Office of Science and Technology Policy. https://www.whitehouse.gov/ostp/ai-bill-of-rights/

[99] Government Accountability Office. 2021. *USDA MARKET FACILITATION PROGRAM: Stronger Adherence to Quality Guidelines Would Improve Future Economic Analyses.* Government report GAO-22-468. Government Accountability Office. https://www.gao.gov/assets/gao-22-468.pdf

[100] U. S. Government Accountability Office. [n. d.]. Record Linkage and Privacy: Issues in Creating New Federal Research and Statistical Information. https://www.gao.gov/products/gao-01-126sp

[101] Office for Civil Rights. 2016. *Nondiscrimination in Health Programs and Activities.* Technical Report 81 FR 96. Office for Civil Rights. 31393 pages. https://www.govinfo.gov/content/pkg/FR-2016-05-18/pdf/2016-11458.pdf#page=19

[102] Office of Inspector General. 2022. *Inaccuracies in Medicare's Race and Ethnicity Data Hinder the Ability To Assess Health Disparities.* Data Brief OEI-02-21-00100. Office of Inspector General. 24 pages. https://oig.hhs.gov/oei/reports/OEI-02-21-00100.pdf

[103] Office of Management and Budget. 1995. *Standards for the Classification of Federal Data on Race and Ethnicity.* Federal Register Notice. Office of Management and Budget. https://obamawhitehouse.archives.gov/node/15639

[104] Office of Management and Budget. 1997. *Revisions to the Standards for the Classification of Federal Data on Race and Ethnicity.* Federal Register Notice. https://obamawhitehouse.archives.gov/node/15626

[105] Office of Management and Budget. 2016. *Managing Information as a Strategic Resource.* Circular A-130. https://www.whitehouse.gov/wp-content/uploads/legacy_drupal_files/omb/circulars/A130/a130revised.pdf

[106] Office of the Federal Register and U.S. Government Publishing Office. 2011. Supplemental Nutrition Assistance and Food Distribution Program. https://www.ecfr.gov/current/title-7/subtitle-B/chapter-II/subchapter-C/part-272/section-272.6#p-272.6(g)

[107] Office of the Inspector General. 2020. *Market Facilitation Program— Interim Report.* Audit Report 03601-0003-31 (1). 43 pages. https://www.usda.gov/sites/default/files/audit-reports/03601-0003-31%281%29_FR_508_FOIA_signed.pdf

[108] Amy O'Hara and Carla Medalia. 2018. Data Sharing in the Federal Statistical System: Impediments and Possibilities. *The ANNALS of the American Academy of Political and Social Science* 675 (Jan. 2018), 138–150. https://doi.org/10.1177/0002716217740863

[109] Margaret O'Mara. 2018. The End of Privacy Began in the 1960s. *The New York Times* (Dec. 2018). https://www.nytimes.com/2018/12/05/opinion/google-facebook-privacy.html

[110] Secretary's Advisory Committee on Automated Personal Data Systems. 1973. *Records, Computers and the Rights of Citizens Report of the Secretary's Advisory Committee on Automated Personal Data Systems.* Technical Report. United States Department of Health, Education, and Welfare, https://archive.epic.org/privacy/hew1973report/.

[111] Panel on Confidentiality and Data Access (U.S.), George T. Duncan, Thomas B. Jabine, Virginia A. De Wolf, National Research Council (U.S.), and Social Science Research Council (U.S.) (Eds.). 1993. *Private lives and public policies: confidentiality and accessibility of government statistics.* National Academy Press, Washington, D.C.

[112] Federal Interagency Working Group on Improving Measurement of Sexual Orientation and Gender Identity in Federal Surveys. 2016. *Evaluations of Sexual Orientation and Gender Identity Survey Measures: What Have We Learned?* Technical Report. https://nces.ed.gov/FCSM/pdf/Evaluations_of_SOGI_Questions_20160923.pdf

[113] Karin Orvis. 2022. Reviewing and Revising Standards for Maintaining, Collecting, and Presenting Federal Data on Race and Ethnicity. https://www.whitehouse.gov/omb/briefing-room/2022/06/15/reviewing-and-revising-standards-for-maintaining-collecting-and-presenting-federal-data-on-race-and-ethnicity/

[114] Collin Peterson. 2008. H.R.2419: Food, Conservation, and Energy Act of 2008. http://www.congress.gov/ Archive Location: 05/22/2008.

[115] Tom Petska. 1997. *Partnerships in Data Sharing: The Internal Revenue Service and the Bureau of Economic Analysis.* Special Contributed Panel. 1997 Joint Statistical Meetings, Anaheim. https://www.irs.gov/pub/irs-soi/datashar.pdf

[116] Neomi Rao. 2017. *EEO-1 Form; Review and Stay.* Memorandum. Office of Management and Budget. https://www.reginfo.gov/public/jsp/Utilities/Review_and_Stay_Memo_for_EEOC.pdf

[117] Madison St. Clair Record. 2022. Davis Introduces Legislation to Block Biden CFPB From Over-regulating the Farm Credit System. *Madison - St. Clair Record* (March 2022). https://madisonrecord.com/stories/626465055-davis-introduces-legislation-to-block-biden-cfpb-from-over-regulating-the-farm-credit-system

[118] Priscilla M. Regan. 1995. *Legislating privacy: technology, social values, and public policy.* University of North Carolina Press, Chapel Hill.

[119] Frank Rep. Pallone. 2022. H.R.8152 - 117th Congress (2021-2022): American Data Privacy and Protection Act. http://www.congress.gov/ Archive Location: 12/30/2022.

[120] Kit T. Rodolfa, Hemank Lamba, and Rayid Ghani. 2021. Empirical observation of negligible fairness–accuracy trade-offs in machine learning for public policy. *Nature Machine Intelligence* 3, 10 (Oct. 2021), 896–904. https://doi.org/10.1038/s42256-021-00396-x Number: 10 Publisher: Nature Publishing Group.

[121] Chris Mills Rodrigo. 2021. Wyden-Paul bill would close loophole allowing feds to collect private data. *The Hill* (April 2021). https://thehill.com/policy/technology/549468-wyden-paul-bill-would-close-loophole-allowing-feds-to-collect-private-data/

[122] Nathan Rosenberg and Bryce Stucki. 2019. How USDA distorted data to conceal decades of discrimination against Black farmers. *The Counter* (June 2019). https://thecounter.org/usda-black-farmers-discrimination-tom-vilsack-reparations-civil-rights/

[123] Bobby Rush. 2021. Farm Subsidy Transparency Act of 2021. https://www.congress.gov/bill/117th-congress/house-bill/3794

[124] Tim Ryan. 2022. The CHIPS and Science Act of 2022. https://www.congress.gov/117/plaws/publ167/PLAW-117publ167.pdf

[125] Alexis R. Santos-Lozada, Jeffrey T. Howard, and Ashton M. Verdery. 2020. How differential privacy will affect our understanding of health disparities in the United States. *Proceedings of the National Academy of Sciences* 117, 24 (June 2020), 13405–13412. https://doi.org/10.1073/pnas.2003714117 Company: National Academy of Sciences Distributor: National Academy of Sciences Institution: National Academy of Sciences Label: National Academy of Sciences Publisher: Proceedings of the National Academy of Sciences.

[126] Charles Schumer. 2021. United States Innovation and Competition Act of 2021. https://www.congress.gov/bill/117th-congress/senate-bill/1260

[127] Charles G Scott. 1999. Identifying the race or ethnicity of SSI recipients. *Soc. Sec. Bull.* 62 (1999), 9.

[128] Carolyn Shettle and Geraldine Mooney. [n. d.]. Monetary Incentives in U.S. Government Surveys. ([n. d.]).

[129] Paige Smith. 2019. EEOC's Pay Data Collection Reinstated by Federal Judge. *Bloomberg Law* (March 2019). https://news.bloomberglaw.com/daily-labor-report/eeocs-pay-data-collection-reinstated-by-federal-judge

[130] Mark E. Stallion. 2021. The bi-partisan IDEA Act: a great idea, or pointless data gathering? *Reuters* (Aug. 2021). https://www.reuters.com/legal/legalindustry/bi-partisan-idea-act-great-idea-or-pointless-data-gathering-2021-08-06/

[131] Ashlie D. Stevens. 2021. The USDA has discriminated against Black farmers for years. Can this legislation bring about change? *Salon* (June 2021). https://www.salon.com/2021/06/10/the-usda-has-discriminated-against-black-farmers-for-years-can-this-legislation-bring-about-change/ Section: Food.

[132] Peter L Strauss. 1984. The place of agencies in government: Separation of powers and the fourth branch. *Columbia Law Review* 84, 3 (1984), 573–669.

[133] Sustainable Agriculture and Food System Funders. 2021. *Comment from Sustainable Agriculture and Food System Funders.* Technical Report USDA-2021-0006-0378. https://www.regulations.gov/comment/USDA-2021-0006-0378

[134] The Leadership Conference on Civil and Human Rights. 2021. *Comment from The Leadership Conference on Civil and Human Rights.* Technical Report.

[135] United States Patent and Trademark Office. 2015. *Memorandum on the Study of Diversity Among Patent Applicants.* Technical Report. https://www.uspto.gov/sites/default/files/documents/Determination%20on%20Diversity%20of%20Applicants.pdf

[136] U.S. Department of Veterans Affairs. 2022. *Equity Action Plan Summary: U.S. Department of Veterans Affairs.* Technical Report. https://www.whitehouse.gov/wp-content/uploads/2022/04/VA-EO13985-equity-summary.pdf

[137] USA Spending. [n. d.].

[138] USDA Farm Service Agency 2016. *Farm Service Agency Programs.* Factsheet. USDA Farm Service Agency. https://www.fsa.usda.gov/Assets/USDA-FSA-Public/usdafiles/FactSheets/2016/farm_service_agency_programs.pdf

[139] Saurabh Vishnubhakat. 2021. Rethinking USPTO Applicant Diversity. *IPWatchdog.com* (Jan. 2021). https://ipwatchdog.com/2021/01/31/rethinking-uspto-

applicant-diversity/id=129505/ Section: IPWatchdog Articles.
[140] Hansi Lo Wang. 2018. Citizenship Question May Be 'Major Barrier' To 2020 Census Participation. *NPR* (Nov. 2018). https://www.npr.org/2018/11/01/663061835/citizenship-question-may-be-major-barrier-to-2020-census-participation
[141] Samuel D. Warren and Louis D. Brandeis. 1890. The Right to Privacy. *Harvard Law Review* 4, 5 (1890), 193–220. https://doi.org/10.2307/1321160 Publisher: The Harvard Law Review Association.
[142] Evan Weinberger. 2022. CFPB Gets March Deadline for Small Business Lending Data Rule. *Bloomberg Law* (June 2022). https://news.bloomberglaw.com/banking-law/cfpb-gets-march-deadline-for-small-business-lending-data-rule
[143] Evan Weinberger. 2022. Regulators Walk Tightrope on Race in Community Lending Update. *Bloomberg Law* (May 2022). https://news.bloomberglaw.com/banking-law/regulators-walk-tightrope-on-race-in-community-lending-update
[144] Gus Wezerek and David Van Riper. 2020. Changes to the Census Could Make Small Towns Disappear. *The New York Times* (Feb. 2020). https://www.nytimes.com/interactive/2020/02/06/opinion/census-algorithm-privacy.html
[145] Orice Williams. 2008. *Race and Gender Data Are Limited for Nonmortgage Lending*. Technical Report GAO-08-1023T. Government Accountability Office. 21 pages. https://www.gao.gov/assets/gao-08-1023t.pdf
[146] Alice Xiang. 2022. Being 'Seen' vs. 'Mis-Seen': Tensions between Privacy and Fairness in Computer Vision. https://doi.org/10.2139/ssrn.4068921
[147] Alice Xiang. 2022. Being'Seen'vs.'Mis-Seen': Tensions between Privacy and Fairness in Computer Vision. *Harvard Journal of Law & Technology, Forthcoming* (2022).
[148] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, and Krishna P. Gummadi. 2017. Fairness Beyond Disparate Treatment & Disparate Impact: Learning Classification without Disparate Mistreatment. In *Proceedings of the 26th International Conference on World Wide Web*. International World Wide Web Conferences Steering Committee, Perth Australia, 1171–1180. https://doi.org/10.1145/3038912.3052660

A APPENDIX: EQUITY ACTION PLAN CONTENT ANALYSIS METHOD

Our content analysis aimed to (a) identify any recognition of barriers to demographic data to enable disparity assessments, and (b) assess the concreteness of plans to overcome such data deficits. To conduct it we read through each of the equity action plans in their entirety. In Table 3, we organize these plans into five main categories: sample surveys, form collection, record linkage, imputation, and visual observation. We marked a category as green if the agency's equity action plan proposed a concrete program that fit into one of these categories, even if it was a pilot. We marked a category as yellow if a partial plan or generic mention appeared, but no concrete plan.

Because EO 13,985 centers on disparities in the general population, we focused on plans to overcome data deficiencies with regard to the public broadly. We thus omitted efforts to collect or compile demographic data about specific subgroups, such as contractors. Additionally, the program had to be substantial, allocating resources to a sizable population.

Our approach captures most, if not all, of agencies' efforts to overcome data deficits, because it is unlikely that agencies are engaged in efforts that are not discussed in these plans. The plans submitted to the White House are meant to be comprehensive, and agencies would want to receive "credit" for their efforts to comply with EO 13,985.

B APPENDIX: DETAILED CASE STUDIES

B.1 Farm Subsidies (FCA, USDA)

There are three major categories of agricultural lenders: a network of public lenders called the Farm Credit System (FCS), which is regulated by the Farm Credit Administration (FCA); commercial lenders; and the USDA's Farm Services Agency (FSA) [56]. Eighty-three percent of agricultural loans are issued through the Farm Credit System or commercial lenders; until now, ECOA prohibited these lenders from asking borrowers for their race, but new rulemaking from the CFPB on ECOA requires the FCA to collect demographic data [21]. The FSA's lending is more specialized, intended to "provide credit to agricultural producers who are unable to receive private, commercial credit, including special emphasis on beginning, minority and women farmers and ranchers"[138], and the agency is required to collect and publicly release demographic data [56].

Given that numerous scholars have documented concerns about racial discrimination in the administration of farm loans and many lawsuits have alleged discrimination against minority farmers [33, 35], the USDA faces particularly strong calls to improve oversight of its programs, be transparent about its resource allocation, and better understand how to serve minority farmers' credit needs. John Boyd, the President and Founder of National Black Farmers Association stated, "until we have full transparency, we can't see the full extent to which USDA programs continue to perpetuate the agency's long history of racism" [131]. Unfortunately, as a 2019 GAO report documents, the lack of demographic data creates challenges for regulators, researchers, and advocates who seek to identify risks of discrimination and enforce fair lending laws [56].

The USDA's current approach to data collection appears insufficient. Although the FSA is not just permitted, but *required* to collect demographic data,[6] its demographic reporting is inconsistent and unreliable. Public comment from Sustainable Agriculture and Food System Funders in 2021 called the website reporting USDA demographic data "very badly out of date, cumbersome and not user friendly" and noted that "much of the data is missing" [133]. Additionally, some of the FSA's race and ethnicity data has historically been collected based on visual observation, even though the agency has been trying to move away from employee-observed data since 2004 due to accuracy concerns. For example, the data management system for the 2018 Market Facilitation Program (MFP), a $14 billion cash assistance program to offset tariff-induced losses, required FSA employees to complete the race and ethnicity field to proceed to the next screen, so employees entered a guess based on visual observation even if applicants declined to self-identify [107]. As a result, over two-thirds of race and ethnicity data was based on visual observation even though USDA analysts agree that visual observation yields unreliable data [55]. The USDA formed a task force in 2021 to improve data management policies and implement a decade-old regulation prohibiting visual observation, but has not planned for alternate sources of demographic data to meet its statutory mandate [107]. The most comprehensive source of demographic data is the Census of Agriculture, conducted every five years, but the Census does not link demographic data to specific loan applications or subsidy programs, making it ill-suited to assessing disparities in specific programs [122].

In 2021, Sen. Cory Booker (D-NJ) and Rep. Bobby Rush (D-IL) introduced the Farm Subsidy Transparency Act, which would require the USDA to track and publicly release the race, gender, and amount

[6] The 2008 Farm Bill required the Secretary of Agriculture to report the race, ethnicity, and gender of applicants and recipients for all FSA programs every other year [114].

received for every individual who applies for assistance from the agency [94]. The legislation amends ECOA, which may also allow demographic data to be collected for some non-FSA agricultural loans. The Farm Subsidy Transparency Act has not been debated since its introduction [123].

Another avenue to enable demographic data collection is Section 1071 of the 2010 Dodd-Frank Act, which requires the CFPB to collect demographic data for small businesses to identify and address violations of ECOA [142]. The CFPB "believes that covering agricultural credit . . . is important" to implementing Section 1071 and included Farm Credit System loans for data collection in its final rulemaking in March 2023 [21]. In response to the CFPB, a bipartisan group of congress members introduced the Farm Credit Administration Independent Authority Act, which prevents the CFPB from issuing policies that affect the FCA, citing concerns that "Farm Credit lenders and borrowers will be subject to excessive, duplicative, and unnecessary reporting requirements that, at the end of the day, will demand new, costly IT infrastructure, additional staff, and will ultimately expect lenders to guess the demographic information of a borrower in the name of 'fair lending' if this is left unreported" [117].

GAO, the USDA, other federal actors, and advocates have all emphasized the essential role of demographic data in bringing transparency to USDA programs, building an understanding of minority farmers' needs, and enabling the USDA to assess its policies. A combination of legal limits, technical and procedural missteps, resource constraints, and, perhaps, political will appear to have stalled robust data collection.

B.2 Food Security (USDA)

Data collection for food security programs like the Supplemental Nutrition Assistance Program (SNAP), Child and Adult Care Food Program (CACFP), and Summer Food Service Program (SFSP) is similar to data collection for farm subsidies: civil rights law requires the USDA to collect demographic data on beneficiaries [3], but as the agency moves away from visual observation, the best mode of data collection is unclear [74].

DOJ regulations require federal agencies to collect racial and ethnic data from program applicants to enforce Title VI; SNAP implements this regulation by requiring state agencies to collect demographic data [106]. As applicants are not required to self-identify, until recently, employees would use visual observation to complete missing data [107]. After reliability concerns and reports that program applicants did not want to have their race determined visually [74], the FNS instituted a phase out period from mid-2021 until the end of 2022 for local agencies to replace visual observation for CACFP and SFSP [1], and proposed removing visual observation for SNAP in June 2022 [53]. Food security-related nonprofits and beneficiary advocates support the phasing out of visual observation. The Massachusetts Law Reform Institute, for instance, argues that visual observation is "rife with bias and discrimination, and results in data collection that is likely error prone, inaccurate, and misleading" in their response to the USDA's request for comments on advancing racial justice and equity [80].

The USDA does acknowledge the end of visual observation could threaten comprehensive collection of demographic data. USDA raised two possible approaches to address this. First, USDA encouraged data collectors to explain the importance of demographic data and encourage participants to self-identify. Second, USDA suggested looking for other sources of demographic data to link records, specifically noting school enrollment records as a possibility for CACFP and SFSP [1]. Since the administration of food security programs is decentralized, different localities will likely employ different strategies to collect demographic data with varying levels of success.

The FNS, participants, and advocates agree that visual observation is a flawed approach, but the best alternative for meeting the agency's data needs is unclear. As with data collection for farm subsidies, the agency does not appear to consider requiring applicants to self-identify, likely due to the risk of deterring applications. Data sharing is presented as the most viable method, but given the decentralized nature of school records, success may vary from place to place.

B.3 Lending

In its report, *Fair Lending: Race and Gender Data Are Limited for Non-Mortgage Lending*, GAO discussed both the merits and costs of requiring lenders to collect sensitive data from non-mortgage loan applicants. While collecting such data "could help address current data limitations that complicate efforts to better assess possible discrimination," GAO also warned of the potential "additional costs on lenders that could be partially passed on to borrowers," which could arise from "information system integration, software development, data storage and verification, and employee training" [145]. Providing a similar view on the significant burdens associated with sensitive data collection, a joint statement from ABA, BPI, CBA, HPC, and MBA suggested that "the challenge of collecting this information, from a practical level . . . largely defeats any corresponding benefit that the Bureau could have in collecting this information" [8]. Inconsistent reporting and interpretation of data, the costs to pursue systemic changes, and a lack of public understanding of its purpose all pose serious challenges to proposed expansions of sensitive data collection both in terms of its ability to garner industry support and the financial feasibility of its implementation.

Potential costs notwithstanding, GAO emphasized that sensitive data could assist in fighting discrimination if collection was mandated. Voluntarily collected data, on the other hand, "would not likely materially benefit efforts to better understand possible discrimination" due to inconsistencies in data collection and the risk that "few lenders would participate out of concern for additional regulatory scrutiny of their non-mortgage lending practices and the potential for litigation" [145]. In the aforementioned joint letter, several industry trade groups also alleged that implementing additional fields for race and ethnicity data under HMDA would "provide limited benefit both from a data integrity and data analysis perspective" [8].

While some have focused on the consequences of collecting too little demographic data, others have warned that proposed changes to CFPB data policies will provide too much data to too many people. In 2014, the CFPB proposed to amend Regulation C, which implements HMDA, to "add several new reporting requirements" and expand its coverage [30]. A 2017 article on increasing data

points to be collected and reported under HMDA argued that data privacy issues had been "largely overlooked" and contentious, as covered entities have questioned why the rule "failed to establish a method to mask certain data fields that would protect an applicant's identity" [89]. The threat of re-identification has influenced approaches to the privacy-bias tradeoff, with some contending that the risk is cause for scrapping proposed expansions of sensitive data collection entirely while others see it as fodder for their case against disaggregation [8].

B.4 Tax administration (IRS)

The Treasury Department and IRS are engaged in a review of racial inequity in tax policy and administration, such as assessing whether the pandemic direct assistance program of Economic Impact Payments (EIPs) was distributed equitably [5]. However, the IRS only collects the information required to uphold the tax code, which excludes race and ethnicity.

The primary approach relies on imputation, using methods based on Bayesian Improved Surname Geocoding (BISG) where filers' surname and location predict their race [34, 41, 58]. While imputation provides individual-level data without any additional data collection, and thus, no legal or procedural negotiation, it does come with limitations. Imputed race has higher error rates than collecting self-reported race, particularly for mixed race and indigenous populations [64] and disparity estimates may be biased [41, 58].

As an alternative, the IRS data has been linked Census data to obtain self-reported race and ethnicity at the individual level. Such IRS-Census linkages are only on a project-specific basis, as such research agreements require detailed, resource-intensive reviews [58]. While privacy law enumerates situations where disclosure is permissible, addressing bias is not explicitly acknowledged as a valid exception.[7] Specific statutes for each agency and program create additional uncertainty about the operative restrictions on combined datasets [90]. If bias assessment were included in an agency's statutory mandate or listed as a potential use when data is collected, data sharing would be better supported by privacy law. The GAO suggests that Congress modify Title 13 to allow the Census to share data with OTA, modeled after existing provisions that allow data exchange with BEA and BLS [58].

B.5 Patents (USPTO)

In 2021, a bipartisan group of Congressmembers re-introduced the Inventor Diversity for Economic Advancement (IDEA) Act into the House and Senate following their initial attempt in 2019 [96]. If signed into law, the IDEA Act of 2021 would empower the US Patent and Trademark Office (USPTO) to collect, analyze and report on demographic information—gender, race and military or veteran status—submitted voluntarily by patent and trademark applicants [61]. In a 71-27 vote, the Senate passed IDEA as an amendment to the US Innovation and Competition Act (USICA) in May 2021 [126], but it was not included in the final iteration of USICA in the CHIPS Act [124].

Proponents of USPTO collecting demographic data have argued that the policy would help reverse "decades of underrepresentation for women, minority and low-income patent applicants" [93] by providing non-proxy information needed to better understand and address the patent disparities [96]. However, some opponents to the IDEA Act have questioned the necessity of burdening applicants and the USPTO with sensitive data collection [71]. During the Senate Judiciary Committee hearing that passed the IDEA Act, Senator John Kennedy (R-LA) questioned the need to collect the data at all when the statistics seemed to already be known [66]. The exact types of demographic data that the USPTO should be able to collect has also been the subject of controversy. The IDEA Act proposes a limited scope of data collection compared to its 2019 version, which included sexual orientation, disability, and age, as some of the additional proposed categories raised concerns [130].

In a 2015 Memorandum on the Study of Diversity Among Patent Applicants, the USPTO discussed the "tension" between a lack of public support for mandatory surveys due to privacy concerns and the method's ability to produce demographic data of better quality and reliability compared to voluntary data collection [135]. The USPTO and other interested parties have observed that public commentary often advocates for voluntary demographic data collection. Indeed, Senator Ted Cruz's (R-TX) pushback against the IDEA Act in the Senate Judiciary Committee centered on his belief that an amendment was necessary to ensure the voluntary nature of any data collection [66]. The Day One Project, an initiative from the Federation of American Scientists is a notable exception to the broader trend towards supporting a voluntary process [139]. In its transition document for the USPTO, published in 2021, the Day One Project listed a pilot program for mandatory demographic data collection among its 25 recommendations [37].

B.6 Hiring (EEOC)

The Equal Employment Opportunity Commission (EEOC) was established under the Civil Rights Act of 1964 to administer antidiscrimination law in the workplace. In service of this mandate, the EEOC requires employers to report the demographic makeup of their workforce annually through survey EEO-1 [84]. After a task force asked the National Academy of Sciences (NAS) to identify the best data strategy to address wage discrimination, the NAS recommended in 2012 that EEO-1 also collect data on hours worked and pay rates [43]. Following feedback from a working group and a pilot study, the Office of Management and Budget (OMB) approved the addition of pay data to the EEO-1 in 2016 for a three-year period [43]. Wage equity advocates hoped that linking demographic and pay data would enable employers to recognize and self-monitor for pay disparities, help the EEOC create statistical tools to flag cases requiring investigation, and arm enforcers with stronger evidence of bias [43, 82].

Federal agencies earned buy-in from employers during this multi-year, deliberative process that included two independent studies, a public hearing with expert testimony, and two rounds of notice and comment [42]. When employer representatives raised concerns about protecting individual privacy in aggregate data releases, the EEOC re-examined statistical confidentiality standards to ensure tables with small cell-counts are kept private. In response to feedback from employer representatives that expanding the EEO-1 would minimally affect administrative costs, the NAS and EEOC opted to

[7] For example, in the Privacy Act [22].

add pay data to EEO-1 in a second component instead of creating a new form [43].

OMB abruptly ended pay data collection in 2017, stating that Component 2 "lacks practical utility, is unnecessarily burdensome, and does not adequately address privacy and confidentiality issues" [116]. After pro-worker groups brought suit against the OMB and EEOC, a federal judge reinstated Component 2 in 2019, finding the OMB's decision arbitrary and capricious and reiterating the value of pay data for self-monitoring and enforcement purposes [129]. The EEOC still opted not to renew Component 2 after the three-year trial period, stating that the prior estimate of employer burden was ten times too low and the "unproven utility" of pay data did not justify the cost [44].

In 2020, led by a different slate of commissioners, the EEOC asked NAS to revisit the best mode of pay data collection and analyze the one-time data collection from 2017 and 2018 [42]. NAS published their findings in a July 2022 report, where they found serious cases of pay inequity: one employer had a −51.3% pay gap for Black men compared to white men [85, p. 214, 216-217], another had a −52.3% pay gap for Hispanic women relative to white women[85, p. 214, 216-217], and unnamed Silicon Valley tech firms had "extreme pay gaps based on race, sex, and/or ethnicity" [42]. NAS affirmed that the lack of pay data has been a longstanding obstacle to enforcing pay discrimination laws [85, p. 14-18] and that collecting this data is essential to assessing pay disparities by sex, race, and ethnicity [85, p. 28]. NAS recommended not just that EEOC reinstate data collection, but that it widen the firms it collects data from and collect more granular pay data. Pay data collection will recommence in July 2023 [60].

Received 6 February 2023; Revised May 2023; Accepted 10 May 2023