When to Collect Sensitive Category Data? Working Towards a Participatory Framework for Balancing Privacy and Freedom from Discrimination in Automated Decision Systems in the Public Sector

Version 1.0 of this paper was prepared for TPRC 50 2022, a shorter version covering the first research questions has been submitted as a poster (preliminary study) to CSCW 2022

ANNA LENHART
University of Maryland, alenhart@umd.edu

Automated Decision Systems (ADS) are being used to inform important decisions in government services. Concerns regarding discrimination in ADS have led to the rise of bias mitigation techniques, or data science practices that measure and adjust for disparities based on protected class data. These techniques often require demographic or sensitive category data to both measure discrimination in ADS and process data to mitigate the discriminatory bias. The collection of sensitive category data increases privacy concerns. This preliminary study includes nine semi-structured interviews with data practitioners working in government. The analysis explores what considerations data practitioners in the public sector make when determining best practices for sensitive category data collection and how they engage with the tradeoff between privacy and right to freedom from discrimination. Themes pertinent to social services emerged including the importance of accessibility, reasons for data minimization and the role of reporting structures and historical norms. The analysis also offers considerations for ways that a structured framework and or stakeholder engagement process may inform the collection and use of sensitive category data, including the intersection of these decisions with the right to an explanation and the right to contest a decision.

Keywords: demographic data, sensitive data, public sector, data privacy, fairness, anti-discrimination

1 BACKGROUND & MOTIVATION

Artificial intelligence, machine learning and rules-based algorithms, a class of data science tools referred to as Automated Decision Systems (ADS) throughout this paper, can be used to make decisions automatically, or influence a human decision. Applications of ADS in the public sector include screening calls related to child protection services (Chouldechova et al, 2018; Eubanks, 2018) determining whether defendants receive pretrial detention (Corbett-Davies et al, 2017; Kleinberg et al, 2016), collecting taxes or deploying emergency response (Veale, 2018). The use of these applications has led to increased concerns over discrimination as data scientists have underscored that the data used to train ADS may be embedded with historical patterns of discrimination (Ajunwa, 2020; Favaretto, 2019) and/or does not represent the environment in which the ADS will operate (Buolamwini & Gebru, 2018; d'Alessandro, 2017).

Concerns regarding discrimination have led to the rise of awareness-based techniques for bias mitigation or data science practices that measure and adjust for disparities based on protected class data (Bellamy et al, 2018). Bias mitigation techniques often require sensitive category data such as race, ethnicity, gender, sexual orientation, age, etc to both measure discrimination in ADS and adjust the computer code and data processing to mitigate the discriminatory bias (Bogen et al, 2020; Zehlike et al, 2017; Zliobaite & Custers, 2016). Collection of sensitive category data puts data subjects at an increased risk of being re-identified, targeted or having records compromised (Bogen, et al, 2020; Madden, 2017).

Legal scholars suggest that disparate impact tests may be required for holding ADS accountable (Hoffman & Podgurski, 2020; Barocas and Selbest, 2016). Williams et al (2018) argues that big data techniques make it difficult to keep aspects of one's identity such as race, gender, socio-economic status (SES), etc private because they can often be inferred, suggesting more collection of structured sensitive category data. Whereas Veale and Binns suggest that discrimination can be mitigated without collecting more sensitive category data (2017).

Recent studies reason that data practitioners should address the tradeoff between privacy and freedom from discrimination by considering context and asking questions about where and how to collect sensitive category data (Andrus, et al 2021; Bogen et al, 2020). As regulatory proposals include mandates to both ensure that ADS are not discriminatory and minimize the collection and storage of private data (Zliobaite & Custers, 2016) it will become more important for ADS developers in the public sector or otherwise to have guidance to navigate choices regarding sensitive category data collection.

Frameworks serve as a useful tool for considering the context surrounding an ADS project (Schwartz et al, 2022; Suresh & Guttag, 2021; Moss et al, 2021; Raji et al, 2020; Reisman, 2018; GAO-21-519SP), for auditing AI auditors (Landers & Behrend, 2022), and for engaging stakeholders and data subjects in algorithm design (Lee et al, 2019). While many frameworks related to the ethics of ADS include mention of sensitive category data, few dive into the particulars of how to balance the privacy risks and responder burden associated with that data collection alongside the role sensitive category data plays in measuring and mitigating bias in ADS. Scholars have considered ways for data subjects to be included in conversations pertaining to algorithm design (Vallejos et al, 2017; Hayes, 2011) and acknowledge the importance of including a range of diverse stakeholders especially when bias is a concern (Crawford, 2016; Rock & Grant, 2016).

This preliminary qualitative study explores two research questions:

RQ 1: How do public sector data practitioners, which tend to have less resources and different incentives and norms compared to the private sector, currently consider tradeoffs between privacy and freedom from discrimination.

RQ 2: How should organizations determine what sensitive category data to collect while considering the data subject's privacy and right to freedom from discrimination? What role (if any) can a framework play? What role (if any) can a stakeholder and data subject engagement play?

2 METHODOLOGY & DATA ANALYSIS

The author conducted nine long form (30-60 minutes) semi-structured interviews with individuals working in data administrator and data analysis roles in the public sector (referred to as data practitioners) over Zoom in the Spring of 2021. Participants were recruited using a snowball method, the author used her network from graduate school and former employment working as a data practitioner in the public sector. This method was preferred because public sector data practitioners are a specialized group generally discouraged from speaking publicly about their work. The project had IRB approval from the University of Maryland (1612009-2), identifiers were removed, and participants were spoken about in general terms to protect their anonymity. The protocol was based on demographic data use considerations within existing AI ethics frameworks (Whittlestone et al, 2019; Algorithm Watch, 2019). The interview was organized into 4 sections described in Table 1.

Table 1: Protocol Outline

Section	Types of Questions
Understanding the Participant	Questions aimed at understanding the participant's roles and responsibilities at the government agencies where they currently or recently served and the details of the data processing projects they contributed to. Examples: • Can you describe your role (current or former) or, in the case of a contractor, a specific project with a governmental body? • What was your relationship like with those who made decisions about when and how to collect data?
Sensitive Category Data Collection	Questions aimed at understanding the types of sensitive category data collected, the manner in which it was collected and how data practitioners think about the data they need and risks to privacy. Examples: • Does your agency follow specific laws/guidelines for the collection (or storage/treatment) of demographic information? • Have you ever chosen not to collect demographic data about individuals in a dataset? If so, why?
Automated Decision Systems	Questions aimed at understanding if and how organizations are using ADS and their ethical considerations. The concepts included: awareness-based techniques for bias mitigation (Bellamy et al, 2018), Proxy variables (Favaretto et al, 2019), ADS audits (Favaretto et al, 2019, Arya et al, 2019), interpretation (explainability) of ADS (Arya et al, 2019, Tomsett et al, 2018), Redress (Almada, 2019), Gaming (Nature, 2016; Veale et al, 2018) • From your point of view what are the hardest questions your team is grappling with regarding ethical/responsible use of ADS? • What is your team's current approach to ADS auditing, if any?
Considerations for the Future	Directed participants to consider future process that may help organizations make decisions regarding when to collect sensitive attribute data. • In your opinion, how should your organization determine what demographic data to collect while considering the data subject's privacy and right to freedom from discrimination? • What type of framework (if any) could be useful? • What types of stakeholder engagement (if any) would be useful?

- What type of data subject engagement (if any) would be useful?

The interview transcripts were analyzed using an iterative coding process with one person, the author, coding. There was no inter-rater reliability or group negotiating. The concepts from literature that informed the protocol also acted as the start list for codes. The first round of coding was done on six of the interviews and focused on getting the participant's responses clustered by the concepts covered in the protocol. After the first round of coding, the code book definitions were updated, and all nine interviews were recoded. The number of participants in each level of government and role is described in Table 2.

Table 2: Participant Description

Category	Example	Number of Participants
Local Government	Data practitioners in city or county governments	5
Federal Government	Data practitioner employed by a federal agency or a government contractor assigned to a federal agency	4
Decision Maker	In a role where the final data collection and processing is determined by the participant	2
Decision Influencer	In an advisory role and able to make suggestions regarding data collection and processing	7

Participants worked on multiple projects ranging from applications for government services, federal procurement, and human resources. Participants emphasized that they can only speak from their perspective, not the organizations for which they worked. Table 3 describes the types of projects discussed during the interview. While participants did describe their projects, their answers to many of the questions in the protocol were informed by their experience on a range of projects. It is worth noting that the author did not speak with anyone working in national defense, criminal justice or a project with risk scoring or facial recognition—projects frequently mentioned in the existing scholarship regarding discrimination in public sector ADS (Corbett-Davies et al, 2017; Eubanks, 2018; Green, 2022).

Table 3: Types of Projects Discussed

Project Type	Decision Systems	Example of Projects	Number of Projects Discussed
Compliance	Rules-based (automated), predictive analytics (automated), Natural Language Processing with rules (automated)	Systems to track a contractor's hours and tasks, or landlord's compliance with affordable housing criteria	4

Information interchange	Rules-based (automated)	Steps in a procurement process, or automatically sending test results	3
Applications for services or funding	Rules-based (automated), human review of raw data (mix of quantitative and qualitative)	Screening for eligibility to receive payments, pensions, funding proposals, services or school admission	4
Program evaluation	No decision system	Randomized control trials, surveys, dashboards to evaluate and justify programs or future programs	5

3 FINDINGS

3.1 Existing privacy and discrimination considerations for data practitioners (RQ 1)

Five themes emerged as participants discussed considerations regarding best practices for sensitive category data collection and how they engage with the tradeoff between privacy and right to freedom from discrimination.

3.1.1 Theme 1: Government is focused on ADS as a means to improve processes

Several participants mentioned a focus on using ADS for improving processes. One working on federal-level projects explained, "[the government is] very conscientious about making decisions about people…there has actually been a focus on prioritizing [ADS] around tasks and processes…but not making suggestions about individuals." Or as one decision maker in local government explains, "I think people get caught up in the shiny stuff and talk about machine learning and data mining and AI as things that they want to do in our organization but there are so many basics we're not even doing as well as we could." Another local government participant described, "we want to automate more data collection and cleaning, data transparency and dashboard processes. But in terms of using data and or using AI [for critical] decision making, I don't think that's something that's emerging anytime soon."

3.1.2 Theme 2: Discrimination concerns centered on accessibility

When discussing discrimination, participant's expressed concerns regarding universal accessibility. One participant, focused on healthcare provision, explained, "People who need the services most are the ones that are less equipped to handle navigating the bureaucracy." Another explained the importance of language when designing user-facing systems, both considerations for non-English speakers and using plain language at a reasonable reading level. Lastly, one participant working on a human resources project expressed concerns that users from an older demographic may be less comfortable with the technical system and may be recorded as "out of compliance" when in actuality they struggled to use the application.

3.1.3 Theme 3: Data minimization was emphasized as a norm for reasons beyond protecting an individual's privacy

When discussing how data is collected and processed, every participant expressed the importance of data minimization. While many mentioned laws such as Confidential Information Protection and Statistical Efficiency Act (CIPSEA), Health Insurance Portability and Accountability Act (HIPAA), and Family Educational Rights and Privacy Act (FERPA), their reasons went beyond legal compliance. A few emphasized the importance of trust in government, one stated, "If we were to allow that data [records on employees] to be compromised, that would irreparably damage our ability to do our work." Many participants discussed internal practices regarding cell suppression and limiting access to personally identifiable

information, one participant working on a project with a high quantity of sensitive data explained, “it was kind of like a code of honor, I suppose that even though you had access to everything you only use what you really need to answer a question.”

Several practitioners mentioned the importance of not asking for more data than what is needed both to decrease burden on respondents and on the organization’s storage and data management systems. One participant, in reference to programs that require applications, said, “I would prefer that we try to minimize the burden on our respondents, and especially questions that might make them uncomfortable.” This participant went on to bring up an issue important to public service, stating “I think about setting expectations… I try to remind [colleagues] that we don’t want to be over promising in our surveys. And we should really only ask questions that are pertinent to what we can do to improve people’s wellbeing, and not things that are far outside the scope of what we can do.” Similarly, a participant working on a grant application process shared concerns that including questions regarding sensitive category data may communicate that demographic factors would contribute to the outcome of an application, even when it would not. Another participant working on a local level shared that the reliance on spreadsheets added challenges for collecting detailed demographics, “many of our providers don’t have sophisticated data systems so collecting additional and more detailed race categories would quickly multiply the number of columns on their spreadsheet.”

3.1.4 Theme 4: Awareness of and use of proxies is a norm

As data practitioners working in the public sector, participants frequently referenced the fact that the data they were working with was informed by larger social structures, one participant explained it this way: “if systemic discrimination and white supremacy and patriarchy and heteronormativity didn’t exist, then [our programs] wouldn’t need to exist.” Most participants understood that it is important to monitor the way data used to inform decisions could include proxy variables—data set features that are strongly correlated with sensitive category data. Specifically, they noted that names and location can be used to predict race and therefore emphasized that they must be mindful if and when those variables are used to influence a decision. For example, one participant described how their organization uses location data: “East West divide by income and race. And if you have certain zip codes, and people with city zip codes and some small townships, people will take that as a proxy [during data analysis].” A few participants working in public health (a sector where demographic data collection is a norm) were able to use sensitive category data to predict future need for services: “and one of the things that we looked at was race as a proxy for a person's COVID risk because that’s the highest predictor of contracting COVID.”

3.1.5 Theme 5: The amount and type of sensitive category data collected is informed by reporting structures and historical norms

Participants described the way that many data fields were determined by the other agencies or funders to whom they report. This was especially true for data practitioners working in local government that need to report to federal agencies. For example, when asked about how the options in demographic data drop down lists is determined, one data practitioner explained, “we try to align with other big system actors in the area, many of our programs are reporting to multiple funders, so there are guidelines that we’re adhering to.” They also described the need for historical comparison, one participant

explained, “for programs that have existed for a very long time, 30, 40, 50, and 60 years, they set the rules for the data that you need to get approved for some program 60 years ago, and they really haven’t changed.” Two local projects related to tracking COVID-19 collected race in part because it was mandated as one participant explained, “the COVID testing forms also asked for demographic data that gets sent to the state because they're tracking [across] the state.”

3.2 Looking forward: policies and frameworks for balancing tradeoffs (RQ 2)

The final section of the semi structured interviews (Table 1) concluded with a conversation about suggestions for how data practitioners and policy makers should navigate the tradeoffs between data subject privacy and the needs to monitor and mitigate discrimination in automated decisions. Five themes emerged from these conversations.

3.2.1 Theme 1: The US needs regulations and mandated (accountable) internal processes

When asked “how should your organization determine what demographic data to collect while considering the data subject’s privacy and right to freedom from discrimination?” several participants immediately began describing the need for regulatory changes and internal processes. One participant, having worked on federal level HR systems emphasized that data scientists/admin often need demographic data and repeatedly asking for permission can be overwhelming for minority groups, they proposed simple mandates that systems not discriminate suggesting that governments collect the data they need to fulfill the mandate and be held responsible for protecting the data explaining, “as a consumer, I don't want to have to think about it ever. I just want the government to just do it right.” she went on to reference the need for an “industry wide standard for how to collect that demographic data and deidentify it.” And mentioned the Census Bureau’s differential privacy efforts (Machanavajjhala et al, 2017).

Two participants both working at the Federal level suggested that decisions pertaining to when to collect demographic data could potentially be approached in a way that is analogous to existing data security requirements. As one explains: “there are agency processes for a lot of technical solutions, but it's generally purely from a security perspective. And security has a very set definition of things that they're looking for. PII is something they're looking for, but not in the sense of bias or discrimination. Just, it's really more in a sense of how is it stored and how is it protected?” A participant working on the local level made a similar comment about the need for evaluators to weigh in on when to collect sensitive category data: “in a perfect world, there would be a much more robust evaluator system where somebody is providing that internal check and maybe advising on when it's important to collect that sensitive information.”

Mandating systems for individuals to contest a decision were repeatedly mentioned as an option for protecting data subjects from discrimination without necessarily needing reams of sensitive category data. One participant working on federal level service provision explained, “you can't really have these automated processes without appeal processes. You know, it's also important that those appeal processes not be 20 times as long as the original one.” Speaking about a future with more ADS projects, one local level participant mentioned that “they [data subjects] should be able to appeal the decision,” going on to list the other considerations for designing an appeals process: “How fast is it? what kind of documents would you need? How can you make it less burdensome from the person who's appealing?”

One participant, speaking specifically to provision of services at a national level, explained that one way to reduce the need for sensitive category data is to make programs more universal and remove criteria that is tied to protected class data, he describes: "you could easily imagine a policy where everybody is meeting, much simpler criteria with fewer loopholes and exceptions for eligibility." It was interesting to hear the participant discuss a change to the underlying policy instead of a change to data collection process, and while making programs universal would increase access to services for everyone, it does not address discrimination concerns in other contexts where ADS is used.

3.2.2 Theme 2: Decisions pertaining to sensitive category data need to include risk analysis and scenario planning

Every participant mentioned that decisions regarding sensitive category data vary by context. To thoroughly understand the context, the data practitioners in the study described the need to assess risks and consider scenarios in which data subjects may be harmed. One participant, working on the federal level tasked with thinking about a range of federal agencies, described, "it depends on the stakes, if we're talking about healthcare data, the stakes could be very, very serious. If we're talking about housing or some of our basic issues [/needs], the stakes are pretty high… whereas I think there are other very low risk scenarios where collecting demographic information, even if something does go wrong, it's not going to traumatize someone, or it's not going to result in someone not receiving services." One participant discussed the role that collaborating with other local governments could play: "if you're talking concrete tools for people making these decisions, I'd say, simplified case studies, or concrete examples would be really useful" going on to describe the importance of scenario planning, explaining that data practitioners need to: "imagine those different scenarios and futures and put themselves in it and be like, okay, which of these would I rather be dealing with?"

3.2.3 Theme 3: Considerations pertaining to sensitive category data are best made during the ADS design phase

When asked about what type of framework could be useful, several participants listed existing frameworks in security practices or under development (most notable, National Institute Standards and Technology's *Towards a Standard for Identifying and Managing Bias in Artificial Intelligence*) and suggested the types of topics and questions that may be helpful for dissecting a specific context and determining when to collect sensitive category data and the underlying structure of the data.

Participants explained that many of these question are/should be considered during the design phase of the project: "before you even get to the data collection part of it, you have an expected outcome, and being able to track that demographic data to your expected outcome, I think is going to be something that people need to be able to do, because that's going to help understand: do we need this data or not?" A participant who works in the design phase on local data processing projects listed the set of questions she thinks about early on suggesting that these questions inform when to collect sensitive category data: "what are your current processes? What are the outcomes you're trying to achieve with this automated system? If you're collecting data and you want to answer questions with that data, where do you foresee yourself using that data in the future?" Another participant in a similar role suggested the framework should push towards helping answer these types of questions: "what do we think are the most important variables to consider? And why? And, what do we think we should definitely not include and why?" These comments are made in reference to local projects where the

data practitioner is often designing the ADS and the data collection instrument, As mentioned in RQ1, sometimes the data infrastructure and data availability is decades old. None of the interviews in the study included discussion of how the use of historic infrastructure would impact the design process.

Participants underscored the importance of needing a framework for evaluating projects that are internal versus ADS that are provided through a vendor. A participant working at the local level explained that distinguishing between if the ADS project is in-house or bought is the first consideration she would include in a framework, "I think procurement is a huge place for things to go right or wrong." Similarly, a participant that implements city-level projects described the questions she wants cities she works with to consider when deploying third party software: "who in the company that runs that software has access to the raw data? You know, a lot of times a company can have access to the data they want to and can use it if they really wanted to? smaller governments may not know to ask about some of that stuff." Another participant working on the local level mentioned that when her agency considers bringing in 3rd party software, she is mindful of ADA compliance and whether the software service keeps the agency in compliance with ADA. One participant explained that local governments need training on "how to cut through the BS from vendors" and learn to "ask critical questions." They went on to mention the difference in business models, agencies serve the public, but software products have a profit motive and it is important when assessing to remember "somebody is getting rich off of this."

3.2.4 Theme 4: There is a role for traditional stakeholder engagement and user testing

Practitioners expressed a range of views regarding how helpful stakeholder engagement could be for navigating sensitive data collection tradeoffs. One participant working on local projects underscored that different areas of the country have specific "local policy objectives" so otherwise national organizations like Black Lives Matter may "take on a different task than some [in other areas of the country]" expanding, "it's really important that we be mindful of how communities are already making change in their local context." They specifically described how these local contexts can even impact the specific demographics that community wants to be collected. One participant working at the Federal level suggested that stakeholders need to represent and be aware of risk groups and "have sufficient sociological understanding of the harms that communities face, but also sufficient technical knowledge to understand how that data would actually be used."

For study participants that worked on projects in which the data subject and user of the ADS were the same, there was an acknowledgement that user research can help governments determine what to collect: "doing more research into the people that are using your services can help you determine what's necessary to collect and what's frivolous." Another participant, working in a similar context to the participant above, explains, "stepping back from my actual work, I'm always like, we should have more data. And we need to be able to talk about race and disparities by race, if we don't, we're not going to be able to do anything." But goes on the explain that this is not a user centric approach, users "want the shortest path through this to get the money [services] that they need. And then all of that, the desire for data goes out the window."

3.2.5 Theme 5: Transparency and education is needed to engage data subjects

The participants that worked on ADS projects in which the data subject was not the user (example: a case worker makes a decision about an applicant's benefits eligibility, ADS user: case worker, data subject: applicant) emphasized that this type of data subject involvement was beyond the scope of user testing and would require a level of transparency and public education outside the scope of most projects: "So it's definitely important to engage with stakeholders, especially people whose information is being collected. But I think there's a transparency element to it there because at the end of the day, users have different levels of understanding about how government works and how technology works. Even if you just ask them, what are your thoughts on how this data is used? Or what we should collect about you, you may not get an answer that is actually going to help drive a [design] decision." One participant mentioned brining in journalists as a means of ensuring an ADS system that impacts a community can be readily understood, "I think journalists have a great sense of how to make complicated things digestible. And like, that's the key here, this shouldn't be behind a black box, this should be understandable to everybody." On the local level, a data practitioner that works with clients and communities explained it this way: "a lot of people probably don't realize that their data is being collected and what it's being used for. So if you start asking too many questions, and they'll be like, *wait a second*. So I think explaining, here's the data, we do collect on you and here's what we do with it."

4 DISCUSSION

This preliminary study supports existing findings that the availability and norms surrounding sensitive category data collection vary by sector and project (Andrus et al, 2021; Bogen et al, 2020; Veale et al, 2018). Additionally, the themes outlined above highlight that questions surrounding sensitive category data can be tied to existing concepts in the field of data rights, such as the right to explanation and the opportunity to contest an ADS' decision (Algorithm Watch, 2019).

4.1 Preparing for limited availability of sensitive category data

A combination of deep-seated data minimization norms, limited data storage infrastructure and archaic picklist options for sensitive categories suggest that data practitioners working in public services need guidance to determine a) in what cases data practices need to be changed so that more sensitive category data is collected and securely stored and b) when techniques for monitoring discrimination that don't require sensitive category data can and should be considered.

Participants in this study along with the academic literature acknowledge that even when demographic data is missing, sock puppet audits (Sandvig et al, 2014) and qualitative studies (Vaele & Binns, 2017) can expose discrimination in an ADS. One data practitioner working at the local level explained, relevant to her program evaluation projects: "We will do data deep dives and conversation with community, we'll bring back data to them [service providers being evaluated] and say, this is what we're seeing, how would you interpret this? What are the conclusions we could draw, etc." acknowledging that qualitative work provides context about what an ADS may be missing.

As mentioned in the findings, participants in the study acknowledged that data points that act as a proxy for race can both encode discrimination into an ADS but also be used to monitor discrimination in an ADS, even if imperfectly. Relying on other studies of discrimination in the environment in which the data represents may help data practitioners identify proxies which can be used to monitor discrimination in the absence of sensitive category data (Vaele & Binns, 2017).

4.2 Reliance on explainability for reducing discrimination and engaging stakeholders

Nearly every ADS discussed in this study was rules based, the only mention of advanced machine learning techniques were projects using natural language processing to pull information out of long text boxes in order to have more information about the program, not for the purpose (at least at the time of the interview) of making a decision about an individual. The data practitioners in this study by and large showed no indication that they foresee more advanced machine learning techniques being used to make critical decisions in their work. The notion that the participants felt confident in their understanding of how ADS made decisions, made them less concerned about discriminatory bias. This is in line with Veale et al's study in which most public sector data scientists used shallow models and underscores that public sector data practitioners value the importance of understanding how an ADS system works and being able to explain it to the public, journalists and other stakeholders (2018).

Throughout the conversations on stakeholder and data subject engagement, participants underlined that being able to explain how the ADS works would be crucial to engaging people in a conversation regarding which demographic data to collect.

One challenge with too much explainability is gaming—or the idea that if a user knows how an ADS works, they may enter untrue data as a means to get their desired outcome (Nature, 2016). Throughout the study, when gaming was mentioned, it was often met with laughter in part because the provision of public services has a long history of being gamed. Two practitioners working in healthcare and education discussed entire industries of lawyers that are hired to help the public "navigate" the bureaucratic and data informed systems that determine who gets benefits or gets into coveted schools, expressing a sentiment that people with means already can "game the system" and some transparency may in fact level the playing field. Similarly, two data practitioners discussed how their screening tools are designed to save users from filling out longer paperwork when they simply don't qualify, so while users often do discover how to answer the screening questions to qualify for the longer application, they are ultimately rejected anyway. Had the study included practitioners working on fraud detection or crime determination, concerns regarding ADS transparency and gaming may have been more pronounced (Veale et al, 2018).

While the participants highlighted that they are not focused on advanced machine learning for critical decisions, they expressed concerns about their agencies bringing in software services that may make decisions in ways that are less transparent to both the public and public servants. This theme speaks to the needed work of scholars like Sloane et al who study the challenges with government procurement of ADS and advocate for specific actions governments can take when procuring ADS including process changes, meaningful transparency, interagency communication and hiring internal experts who can critically probe proposed ADS projects (Sloane et al, 2021; Sloane et al, 2022)

4.3 Reliance on contestability processes

As noted in the findings, participants were adamant that all of the ADS projects they have worked on included the ability for the public to contest decisions, in many ways this is just part of bureaucratic processes – when you get a notice from

the government, at a bare minimum you can "email a person at the city." Participants, however, also acknowledged that these systems are often not well designed. Two participants working in government contracting roles explained that if a user contested a decision the agency would have to flag the issue for them, adding an additional step in the contestability process. This sentiment is echoed in the academic literature on contestability of algorithmic decisions (Lyons et al, 2021; Hirsch et al 2017; Almada, 2019). Additionally, information asymmetries resulting from limited transparency into ADS can make it harder for individuals to contest decisions (Bayamlioglu, 2018). To the extent that public sector data practitioners struggle to ongoingly monitor discrimination in ADS using sensitive category data, *designing for contestability* (Almada, 2019) will only become more critical.

5 FUTURE WORK & FRAMEWORK CONSIDERATIONS

The next step in this line of inquiry includes creating and testing a framework for engaging users, data subjects and key stakeholders in sensitive category data collection policy formulation focused on ADS in social services.

Question for a framework to consider:

- Under what scenarios will supplements and replacements for incomplete sensitive category data be permitted within an ADS project? When instead will an ADS project have to be terminated or paused until better sensitive category data is available?
- How should historical data be treated when practices and structures pertaining to sensitive category data collection change?
- When is gaming in public services a legitimate concern? If explanations have to be concealed due to gaming, what additional oversight (or sensitive category data for internal monitoring) may be needed to account for the fact that the public will have less ability to flag potential discrimination?
- How should the availability of sensitive category data be weighed during government procurement of ADS?
- How should the quality and accessibility of processes for contesting ADS outcomes factor into the need for additional oversight (or sensitive category data for internal monitoring)?
- How much do stakeholders, including data subjects need to understand about an ADS to contribute to decisions regarding sensitive category data collection and treatment?

6 ACKNOWLEDGEMENTS

The author wishes to thank: the study participants who are determined to make government work better for all of us and are on the front lines thinking critically about the role of data in public services, her advisor Dr. Katie Shilton, the University of Maryland iSchool community, scholars at 2020 EASST/4S Conference *Locating and Timing Matters: Significance and agency of STS in emerging worlds* who provided feedback on the background to this study.

REFERENCES

Ajunwa, I. (2020). The “black box” at work. *Big Data & Society*, *7*(2), 2053951720966181.

Algorithm Watch. (2019). AI Ethics Guidelines Global Inventory. Retrieved on Dec, 2, 2019: https://algorithmwatch.org/en/project/ai-ethics-guidelines-global-inventory/

Almada, M. (2019, June). Human intervention in automated decision-making: Toward the construction of contestable systems. In *Proceedings of the Seventeenth International Conference on Artificial Intelligence and Law* (pp. 2-11).

Andrus, M., Spitzer, E., Brown, J., & Xiang, A. (2021, March). 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* (pp. 249-260).

Arya, V., Bellamy, R. K., Chen, P. Y., Dhurandhar, A., Hind, M., Hoffman, S. C., ... & Zhang, Y. (2019). One explanation does not fit all: A toolkit and taxonomy of ai explainability techniques. *arXiv preprint arXiv:1909.03012*.

Barocas, S., & Selbst, A. D. (2016). Big data's disparate impact. *California law review*, 671-732.

Bayamlioglu, E. (2018). Contesting automated decisions. *Eur. Data Prot. L. Rev.*, *4*, 433.

Bellamy, R. K., Dey, K., Hind, M., Hoffman, S. C., Houde, S., Kannan, K., ... & Zhang, Y. (2019). AI Fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias. IBM Journal of Research and Development, 63(4/5), 4-1.

Bogen, M., Rieke, A., & Ahmed, S. (2020, January). Awareness in practice: tensions in access to sensitive attribute data for antidiscrimination. In *Proceedings of the 2020 conference on fairness, accountability, and transparency* (pp. 492-500).

Buolamwini, J., & Gebru, T. (2018, January). Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency* (pp. 77-91). PMLR.

Chouldechova, A., Benavides-Prado, D., Fialko, O., & Vaithianathan, R. (2018, January). A case study of algorithm-assisted decision making in child maltreatment hotline screening decisions. In *Conference on Fairness, Accountability and Transparency* (pp. 134-148). PMLR.

Corbett-Davies, S., Pierson, E., Feller, A., Goel, S., & Huq, A. (2017, August). Algorithmic decision making and the cost of fairness. In *Proceedings of the 23rd acm sigkdd international conference on knowledge discovery and data mining* (pp. 797-806).

Crawford, K. (2016). Artificial intelligence’s white guy problem. *The New York Times*, *25*(06).

d'Alessandro, B., O'Neil, C., & LaGatta, T. (2017). Conscientious classification: A data scientist's guide to discrimination-aware classification. *Big data*, *5*(2), 120-134.

Eubanks, V. (2018). *Automating inequality: How high-tech tools profile, police, and punish the poor*. St. Martin's Press.

Favaretto, M., De Clercq, E., & Elger, B. S. (2019). Big Data and discrimination: perils, promises and solutions. A systematic review. *Journal of Big Data*, *6*(1), 1-27.

Green, B. (2022). The flaws of policies requiring human oversight of government algorithms. *Computer Law & Security Review*, *45*, 105681.

Hayes, G. R. (2011). The relationship of action research to human-computer interaction. *ACM Transactions on Computer-Human Interaction (TOCHI)*, *18*(3), 1-20.

Hirsch, T., Merced, K., Narayanan, S., Imel, Z. E., & Atkins, D. C. (2017, June). Designing contestability: Interaction design, machine learning, and mental health. In *Proceedings of the 2017 Conference on Designing Interactive Systems* (pp. 95-99).

Hoffman, S., & Podgurski, A. (2020). Big Bad Data: Law, Public Health, and Biomedical Databases. In *The Ethical Challenges of Emerging Medical Technologies* (pp. 229-233). Routledge.

Kleinberg, J., Mullainathan, S., & Raghavan, M. (2016). Inherent trade-offs in the fair determination of risk scores. *arXiv preprint arXiv:1609.05807*.

Landers, R. N., & Behrend, T. S. (2022). Auditing the AI auditors: A framework for evaluating fairness and bias in high stakes AI predictive models. *American Psychologist*.

Lee, M. K., Kusbit, D., Kahng, A., Kim, J. T., Yuan, X., Chan, A., ... & Procaccia, A. D. (2019). WeBuildAI: Participatory framework for algorithmic governance. *Proceedings of the ACM on Human-Computer Interaction*, *3*(CSCW), 1-35.

Lyons, H., Velloso, E., & Miller, T. (2021). Conceptualising contestability: Perspectives on contesting algorithmic decisions. *Proceedings of the ACM on Human-Computer Interaction*, *5*(CSCW1), 1-25.

Machanavajjhala, A., He, X., & Hay, M. (2017, May). Differential privacy in the wild: A tutorial on current practices & open challenges. In *Proceedings of the 2017 ACM International Conference on Management of Data* (pp. 1727-1730).

Madden, M. (2017). Privacy, security, and digital inequality.

Moss, E., Watkins, E. A., Singh, R., Elish, M. C., & Metcalf, J. (2021). Assembling accountability: algorithmic impact assessment for the public interest. *Available at SSRN 3877437*.

Nature, 2016. More accountability for big-data algorithms. *Nature* 537**,** 449 (2016). https://doi.org/10.1038/537449a

Perez Vallejos, E., Koene, A., Portillo, V., Dowthwaite, L., & Cano, M. (2017, June). Young people's policy recommendations on algorithm fairness. In *Proceedings of the 2017 ACM on web science conference* (pp. 247-251).

Raji, I. D., Smart, A., White, R. N., Mitchell, M., Gebru, T., Hutchinson, B., ... & Barnes, P. (2020, January). Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing. In *Proceedings of the 2020 conference on fairness, accountability, and transparency* (pp. 33-44).

Reisman, D., Schultz, J., Crawford, K., & Whittaker, M. (2018). Algorithmic impact assessments: A practical framework for public agency accountability. *AI Now Institute*, 1-22.

Rock, D., & Grant, H. (2016). Why diverse teams are smarter. *Harvard Business Review*, *4*(4), 2-5.

Sandvig, C., Hamilton, K., Karahalios, K., & Langbort, C. (2014). Auditing algorithms: Research methods for detecting discrimination on internet platforms. *Data and discrimination: converting critical concerns into productive inquiry*, *22*, 4349-4357.

Schwartz, R., Vassilev, A., Greene, K., Perine, L., Burt, A., & Hall, P. (2022). Towards a Standard for Identifying and Managing Bias in Artificial Intelligence. *National Institute Standards and Technology*

Sloane, M., Chowdhury, R., Havens, J. C., Lazovich, T., & Rincon Alba, L. (2021). AI and Procurement-A Primer.

Sloane, M., Moss, E., & Chowdhury, R. (2022). A Silicon Valley love triangle: Hiring algorithms, pseudo-science, and the quest for auditability. *Patterns*, *3*(2), 100425.

Suresh, H., & Guttag, J. (2021). A framework for understanding sources of harm throughout the machine learning life cycle. In *Equity and access in algorithms, mechanisms, and optimization* (pp. 1-9).

Tomsett, R., Braines, D., Harborne, D., Preece, A., & Chakraborty, S. (2018). Interpretable to whom? A role-based model for analyzing interpretable machine learning systems. *arXiv preprint arXiv:1806.07552*.

Veale, M., & Binns, R. (2017). Fairer machine learning in the real world: Mitigating discrimination without collecting sensitive data. *Big Data & Society*, *4*(2), 2053951717743530.

Veale, M., Van Kleek, M., & Binns, R. (2018, April). Fairness and accountability design needs for algorithmic support in high-stakes public sector decision-making. In *Proceedings of the 2018 chi conference on human factors in computing systems* (pp. 1-14).

Whittlestone, J., Nyrup, R., Alexandrova, A., & Cave, S. (2019, January). The role and limits of principles in AI ethics: towards a focus on tensions. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society* (pp. 195-200).

Williams, B. A., Brooks, C. F., & Shmargad, Y. (2018). How algorithms discriminate based on data they lack: Challenges, solutions, and policy implications. *Journal of Information Policy*, *8*(1), 78-115.

Zehlike, M., Castillo, C., Bonchi, F., Hajian, S., & Megahed, M. Fairness Measures: Datasets and software for detecting algorithmic discrimination, 2017. *URL http://fairness-measures. org*.

Žliobaitė, I., & Custers, B. (2016). Using sensitive personal data may be necessary for avoiding discrimination in data-driven decision models. *Artificial Intelligence and Law*, *24*(2), 183-201.