Govern With, Not For: Understanding the Stuttering Community's Preferences and Goals for Speech AI Data Governance in the US and China

Jingjin Li[1], Peiyao Liu[2], Rebecca Lietz[2], Ningjing Tang[3], Norman Makoto Su[2], Shaomei Wu[1]

[1]AImpower.org, USA
[2]University of California Santa Cruz, USA
[3]Carnegie Mellon University, USA
jingjin@aimpower.org, pliu62@ucsc.edu, rlietz@ucsc.edu, ningjint@andrew.cmu.edu, normsu@ucsc.edu,
shaomei@aimpower.org

Abstract

Current AI datasets are often created without sufficient governance structures to respect the rights and interests of data contributors, raising significant ethical and safety concerns that disengage marginalized communities from contributing their data. Contesting the historical exclusion of marginalized data contributors and the unique vulnerabilities of speech data, this paper presents a disability-centered, community-led approach to AI data governance. More specifically, we examine the stuttering community's preferences and needs around effective stuttered speech data governance for AI purposes. We present empirical insights from interviews with stuttering advocates and surveys with people who stutter in both the U.S. and China. Our findings highlight shared demands for transparency, proactive and continuous communication, and robust privacy and security measures, despite distinct social contexts around stuttering. Our work offers actionable insights for disability-centered AI data governance.

Introduction

Recent advancements in artificial intelligence – particularly large language models and Generative AI – have heavily relied on vast amounts of human data (Bender et al. 2021). Meanwhile, the rights and use of such data for AI development have become increasingly contested issues among scholars, policymakers, and the public (Leslie et al. 2022; Wong 2023). As the extraction and control of personal data by large corporations further concentrate power within AI systems and broader society (Eubanks 2018), our study seeks to redistribute this power by centering the values and agency of a marginalized community – people who stutter (PWS) – in the exploration and design of ethical data governance structures for stuttered speech data.

A second motivation of this work stems from the heightened safety and privacy risks associated with speech data for the stuttering community. As social stigma and negative stereotypes around stuttering are still pervasive, the risks of contributing one's speech data include potential misrepresentation, leak of medical and health-related data, and involuntary disclosure of vulnerable personal identity, especially when the data is used in contexts without the speaker's approval or oversight. Our work seeks to surface and address these risks through open dialogues with the community, fostering shared understanding and collaborative strategies for risk mitigation. In doing so, we also aim to raise awareness among community stakeholders who may not have fully considered the ethical implications of working with speech data from marginalized communities.

Finally, our work seeks to better understand the geocultural factors in AI data governance. Existing research has shown that privacy norms and trust in institutions vary across locales (Bellman et al. 2004; Smith, Dinev, and Xu 2011), and the stigma around stuttering – although present globally – is significantly stronger in China than in the United States (Ma et al. 2023; Ip et al. 2012). As two leading players in the digital economy, China and the U.S. differ remarkably in both their regulatory approaches to AI and personal data and in their citizens' attitudes towards digital privacy. While current U.S. data protection laws are sector-specific and fragmented (US Privacy and Civil Liberties Oversight Board 2025), the recently enacted China's Personal Information Protection Law (PIPL) (The National People's Congress of PRC 2021) provides broader personal data protection and control to consumers, although it applies only to companies but not the government (Kuzio et al. 2022). Despite limited cross-country data on attitudinal difference in AI data governance, prior research indicated that Chinese citizens are generally more accepting of surveillance technologies – such as face recognition – than their U.S. counterparts, particularly when deployed by the government than private enterprises (Kostka, Steinacker, and Meckel 2021).

By comparing the perspectives of the stuttering community in these two countries, we aim to challenge the Western-centered, one-size-fits-all approaches to ethical data governance, and highlight the complex socio-geopolitical contexts that shape the community's needs and preferences regarding how their data should be used and governed.

Guided by these motivations, our research asks:

- RQ1: How can we use and govern disability related data in an ethical, respectable, and power-sharing way that maximizes the community's agency and control?

- RQ2: How do socio-geographical differences affect the community's preferences and needs with respect to data governance?

Rather than treating PWS as passive data subjects, we en-

gage them as the knowers and experts in shaping a dataset that reflects their needs, experiences, and collective priorities. Through interviews with stuttering advocates and a survey with PWS community members in both China and the United States., we explore community preferences on how stuttered speech data should be collected, used, and governed. Our key findings are: 1) participants across both regions shared clear preferences for how their data should be used, largely aligning with prior literature on disability-centered technology design. Use cases that directly benefit the stuttering community—such as improving ASR accessibility, advancing clinical support, and supporting public advocacy—-were consistently favored over commercial or opaque applications; 2) participants emphasized the need for ethical data governance practices that go beyond one-time consent. They strongly advocated for continuous and proactive updates on how their data is used, as well as robust security and privacy-enhancing mechanisms that offer them ongoing control and protection. However, it is difficult for low-resourced, under-staffed organizations to implement these needs; and 3) despite substantial differences in how stuttering is perceived culturally—such as the heightened stigma and disclosure challenges reported in China—we found a similarity in governance values across both communities. This suggests a shared set of needs and priorities among PWS globally, emphasizing the universal importance of transparency, trust, and agency in the stewardship of disability-related data.

Related Work

Stuttering and Speech AI

Stuttering is a neurodevelopmental condition that affects more than 1% of the world population in behavioral, emotional, and cognitive aspects (Bloodstein, Ratner, and Brundage 2021). In addition, people who stutter are impacted by cultural attitudes and social stigma around speech disfluencies and disability (Ip et al. 2012). Stuttering varies widely—both across and within individuals—ranging from overt disfluencies (e.g., repetitions) to covert behaviors like word substitution (Constantino, Campbell, and Simpson 2022; Tichenor and Yaruss 2021). These inherent diversities make it challenging to develop AI applications that work well for stuttered speech. For example, automatic speech recognition (ASR) systems often have significant problems processing stuttered speech, showing word error rates (WER) high as 50% in cases of severe stuttering, which is 10 times the reported consumer average of 5% (Lea et al. 2023). As a result, speech AI applications present significant barriers for users who stutter, which can lead to not only access issues but also negative psychological effects such as increased self-consciousness and lowered self-esteem (Coalson et al. 2022; Wenzel et al. 2023).

As modern ASR models are often trained over thousands of hours of typical, fluent speech (Radford et al. 2023; Ardila et al. 2020; Kourkounakis, Hajavi, and Etemad 2021), the lack of diverse speech data sets remains a major bottleneck for developing inclusive speech AI models that work well for stuttering (Green et al. 2021; Tobin and Tomanek 2022;

Lea et al. 2023).

Despite increasing attention and efforts on creating atypical speech datasets (Lea et al. 2021; MacDonald et al. 2021; Bernstein Ratner and MacWhinney 2018; Li and Wu 2024), important questions and practical challenges remain about how to collect and govern data from marginalized communities in an ethical manner, such as informed consents, long-term monitoring, collective ownership, and the potential misuse of speech data tied to vulnerable identity. Our work focuses on these ethical considerations around data collection and governance. We advocate for the involvement of people with disabilities in decision-making processes around speech AI development and data stewardship, aiming to create research guidelines that ensure technologies reflect their collective values and lived experiences.

Disability Datasets for AI

Recent progress in AI development has raised a growing concern of the AI fairness and accessibility issues for people with disabilities (PWD) (Guo et al. 2019; Whittaker et al. 2019). The under- and mis-representation of PWD in the data used to train and evaluate AI systems (Guo et al. 2019; Ggottermeier Skushalnagar Raja 2016) has directly contributed to performance disparities observed across various AI applications when interacting with PWD (Whittaker et al. 2019; Gurari et al. 2018; Buyl et al. 2022)–from vision tools mislabeling images taken by blind users (Gurari et al. 2018) to biased resume-screening algorithms (Buyl et al. 2022; Glazko et al. 2024).

Researchers and industry practitioners have attempted to mitigate these disparities through repurposed datasets (e.g., VizWiz (Bigham et al. 2010)) and synthetic disability data generation (Wu et al. 2019; Kourkounakis, Hajavi, and Etemad 2021). However, these approaches often raise issues, including extended consent beyond their original collection purpose or reinforcing stereotypes (Whittaker et al. 2019; Kiger 1992; Silverman, Gwinn, and Van Boven 2015).

Recognizing the limitations of repurposed dataset and synthetic dataset, there has been growing efforts to collect AI data directly from PWD (Kamikubo et al. 2024; Kamikubo, Lee, and Kacorri 2023; Sharma et al. 2023; Park et al. 2021). Theodorou et al. (2021) defined a "disability-first dataset" as one that serves a disability community first before potentially generalizing to serve all people through the innovation it enables. A core aspect of creating disability-first datasets is ensuring that the experiences and perspectives of people with disabilities are represented. However, disability-first datasets remain rare in the AI industry, as current data collection practices – even when focused on disabled populations – are largely commercial and expert-driven, often depriving PWD of the power and agency to make decisions about their data in AI development. As a counterexample, this study adopts a disability-first approach to establish a governance framework for data *of* PWD, *for* PWD, and *with* PWD.

Data Justice

In response to concerns about data extraction and marginalization, data justice has emerged as a critical lens for re-

thinking data governance through the lens of equity, agency, and power redistribution, especially in contexts involving disability, race, and structural exclusion. Data justice has become a key framework for addressing these issues, centering on structural power imbalances inherent in data collection, use, and governance (Taylor 2017; Dencik 2022). Rather than being neutral or objective, data infrastructures often reflect and reproduce longstanding forms of exclusion, especially for communities that have historically been denied control over how they are represented and whose voices are routinely marginalized or misinterpreted (Benjamin 2023; Taylor 2017). Data justice critiques how mainstream data practices often reproduce systemic inequities, which is particularly pronounced in disability contexts, where the biopolitical nature of data (Benjamin 2018) intersects with histories of surveillance, medicalization, and paternalistic control over bodies and voices.

Mainstream dataset practices are frequently extractive, which means that data is taken from individuals or communities without effective consent, and then repurposed for use in systems that those communities neither control nor benefit from (Tuck and Yang 2014; Zuboff 2023). Marginalized contributors are often excluded from these decision-making processes, thereby further reinforcing digital marginality (Costanza-Chock 2020). As Wong (2023) argues, even well-intentioned data collection efforts at data stewardship often operate within opaque institutional norms and rarely deliver on their promise of meaningful community control. This requires rethinking governance as a collaborative process through participatory data stewardship. Some research emphasizes that stewardship must be understood as ongoing care work, not technical oversight (Tran et al. 2022; Tseng et al. 2024). Tuck and Yang (2014) also advocate for a politics of refusal, reminding us that community members have the right not only to participate, but also to withhold their data when systems fail to offer safety, accountability, or shared authorship.

Efforts to center community values as the core of data governance still face many constraints and challenges. As Lin et al. (2024) argue, community-based organizations are often treated as peripheral actors in AI-for-social-good (AI4SG) partnerships despite being best positioned to define local needs and ethical boundaries. And Tseng et al. (2024) emphasize the practical complexity of balancing the burden and benefits of contributors, particularly in the clinical setting, where data contributions may be sensitive, and long-term accountability is critical. Their findings indicate the need to propose justice-centered data stewardship that reimagine community engagement not as a one-time act of consent, but as a sustained relationship grounded in mutual respect and collective decision-making. Recent research expands on these themes with concrete proposals for justice-centered stewardship. Hsieh et al. (2024) advocate for Worker Data Collectives, structures designed to redistribute control over data labor to those who generate it. Similarly, Foley, Sylvain, and Foster (2022) call for "community-governed network commons" as a structural alternative to centralized AI governance. Although growing discussions around participatory AI governance, researchers and practi-

tioners still need to operationalize these principles into technical efforts and practices, prioritizing the autonomy and safety of marginalized communities.

This study represents our effort to practice community-centered, collective decision-making at scale, involving over a hundred community participants across two countries.

Method

Recognizing the stuttered community as the key stakeholder and knowledge bearer, we adopted a mixed-methods approach to explore our research questions. We first conducted semi-structured interviews with eight advocates affiliated with the stuttering community to ground our understanding on the community's values, priorities, and concerns surrounding data governance. These insights then informed our design of a survey with the broader community in China and the U.S. to further unpack their preferences and requirements around the use and governance of stuttered speech data for AI development.

Interviews with Stuttering Advocates

Participants and Recruitment To scope out the discussions and considerations for community-centered stuttered speech data governance, we conducted semi-structured interviews with community representatives and domain experts. Through purposive and snowball sampling, we recruited eight prominent community organizers and stuttering advocates, prioritizing participants with strong community ties and experience in collecting or sharing stuttering-related data. Table 1 summarizes their backgrounds.

Six of the eight participants self-identified as PWS, and all held significant professional or advocacy roles in the stuttering community. Recognizing the vast heterogeneity within the stuttering community, we intentionally sought representation across diverse backgrounds, cultural contexts (the US and China), and areas of expertise, ranging from academia, technology, healthcare, to community organizing.

Data Collection Procedure Interviews were conducted by the first and the last author via Zoom, and lasted between 60 to 90 minutes. With participants' consent, each session was recorded and transcribed using automated speech-to-text software. Each participant received a $50 Amazon e-gift card as compensation. The interview focused on understanding ethical, inclusive, and community-centered approaches to stuttered speech data governance: 1) We started by asking participants to share their personal and professional backgrounds, including their connection to the stuttering community and any previous experiences with data collection or technical design, which allowed them to reflect on the positive and negative aspects of those experiences. 2) Next, we introduced the data collection process of stuttered speech, and asked participants for feedback on how a data collection initiative could benefit—not exploit—the stuttering community. These conversations were informed by a critical stance toward traditional research paradigms that often extract data from marginalized communities without returning value or agency (Paris and Winn 2013). 3) We then asked participants about appropriate models for data

Pseudonym	Primary Role(s)	Background Summary
Monica	Stuttering advocate, stuttering-centered podcaster & filmmaker	PWS; creator of a popular podcast on stuttering; currently producing and directing a documentary exploring the intersection of race and disability.
Alex	Stuttering community organizer, artist	Co-leads a nonprofit dedicated to changing public perceptions of stuttering; background in arts and inclusive design.
Nancy	Stuttering advocate, researcher, educator	PWS; directs a university stuttering lab, researches and teaches stuttering to future clinicians, and organizes community-based support programs.
Eric	Stuttering and LGBTQ+ advocate, Technologist	PWS; works in healthcare IT; leads multiple social and support groups for LGBTQ+ & stuttering voices.
Jane	Stuttering advocate and community organizer, doctoral researcher in SLP	PWS, leads stuttering support and advocacy initiatives in both US and China; co-founder of the largest Chinese stuttering community.
Ray	Stuttering community organizer, AI researcher	PWS, co-founder of the largest Chinese stuttering community, co-led the development of a large Mandarin stuttered speech dataset; works on speech technology in a large tech company in EU.
Terasa	Stuttering community organizer, data scientist	PWS, active member in both U.S. and Chinese stuttering community, co-led the Mandarin stuttered speech collection; works in a large tech company in the U.S.
Natalie	Professor in Hearing and Speech Sciences, co-founder of disfluent speech database	Specialist in stuttering and communication disorders; co-founded and co-manage one of the most widely used open disfluent speech database.

Table 1: Summary of interview participants and their expertise

use and access (e.g., open-source vs. permission-based), acceptable use cases (e.g., research, commercial, or both), and the conditions under which commercial use might be justified. We also discussed community expectations around consent, ownership, oversight, and long-term responsibility for the dataset. 4) We concluded the interview by inviting participants' feedback and suggestion an ask for their willingness to stay involved in reviewing projects related to stuttered speech data collection and governance.

Despite having a diverse set of interview participants, we began to see consistent high-level themes emerge after the first six to seven interviews and concluded this part of the study after the 8th interview.

Data Analysis We used an inductive open-coding approach (Saldaña 2021; Charmaz 2014) to analyze the interview transcripts. Our analysis consists of the following steps: 1) First, the first author reviewed the transcripts of each participant after the interview and generated initial codes by annotating the transcripts and adding comments. For example, we had comments such as "de-identification", "controlled access" to describe the measures of data safeguards. 2) Next, the first author discusses the transcripts with the last author who also conducted the interview. Through this discussion, the first author refined the initial comments into a set of agreed-upon codes, organized these codes into broader categories. For example, codes such as "Ongoing updates", "Data use reports" were group under the "transparency beyond data collection" 3) Then the first author thoroughly reviewed all the transcripts multiple times, applying codes as comments and continuously refining the coding scheme through an iterative process. 4) Finally, the whole team collaboratively identified key thematic insights

emerging from the categorized codes and synthesized these insights for reporting.

Survey with Stuttering Community Members in China and the US

Guided by the insights from the interview study, we designed and distributed an online survey to gather broader inputs from members of the stuttering communities in China and the US. The two contexts offered a contrast in privacy laws, trust, and surveillance, shaping our lens for interpreting similarities and differences in governance values. The survey focused on ethical concerns, data-sharing preferences, and expectations for stuttered speech data use and governance. To explore the potential influence of demographic factors (e.g. age, education) and the stigma of stuttering on data governance, we also collected basic demographics and attitudes towards stuttering in the survey.

Participants and Recruitment We recruited participants who self-identified as people who stutter through stuttering advocacy and support organizations both in China and the US. We included participants across different age groups, gender identities, educational backgrounds, and stuttering severity levels to reflect the diversity of PWS communities in each country. A total of 149 participants completed the survey, with 83 respondents from China and 66 from the US (English-speaking). See detailed breakdown of genders, age groups, and severity of stuttering in Table 2. Participants received 20 RMB or $10 USD for completing each survey, with the option to donate their compensation to a designated community organization. The compensation amount was determined in consultation with our partner community organizations. Each survey took approximately 10–15 minutes.

Category	Chinese (n=83)	U.S, (n=66)
Gender		
Male	49 (59.0%)	44 (66.7%)
Female	34 (41.0%)	21 (31.8%)
Non-binary	–	1 (1.5%)
Age		
Under 18	2 (2.4%)	2 (3.0%)
18–24	16 (19.3%)	8 (12.1%)
25–34	53 (63.9%)	19 (28.8%)
35–44	9 (10.8%)	20 (30.3%)
45+	3 (3.6%)	17 (25.8%)
Stuttering Severity		
Covert	9 (10.8%)	14 (21.2%)
Mild	32 (38.6%)	22 (33.3%)
Moderate	34 (41.0%)	25 (37.9%)
Severe	8 (9.6%)	5 (7.6%)

Table 2: Demographics of survey participants

Survey Questions We designed the survey to collect both demographic and attitudinal data on participants' experiences of stuttering, attitudes toward speech data sharing, privacy concerns, and preferences for governance and participation in stuttered speech datasets. The survey included a mix of Likert-scale, multiple-choice, with several open-ended items inviting elaboration. We developed the English and Simplified Chinese versions of the survey, and the translation was verified by bilingual researchers to ensure cross-linguistic consistency. The survey consists of the following sections: 1) Stuttering background, including questions about stuttering identity, gender, age, education level, city of residence, and professional field, prior experience participating in stuttering speech data collection; 2) Personal experiences and attitudes to stuttering, including perceptions of stuttering severity, frequency of avoidance behaviors, and internalized negative attitudes toward stuttering; 3) Data sharing preferences, including ratings on comfort level of sharing speech data with different types of institutions—ranging from universities and nonprofits to government, corporate, and media entities, acceptability of various purposes for which their data might be used, motivating factors (e.g., compensation, being kept informed, data anonymization); 4) Concerns about the potential risks, such as data misuse, privacy breaches, and the development of discriminatory or harmful generative AI content; 5) Governance and protection mechanisms, including rating the importance of the protection measures such as easy ways to delete data, pre-screening of data users' qualifications, and open communication channels, and regular updates on the data use.

Data Analysis We administered the English survey through Google Forms and the Chinese survey through Tencent Questionnaire. We used Python to conduct descriptive statistics and nonparametric inferential tests to examine cross-geographical patterns in attitudes, preferences, and concerns surrounding stuttered speech data collection and governance. We conducted non-parametric comparisons using the Mann-Whitney U test, which is suitable for ordinal data and does not assume normal distribution of responses (Jamieson 2004).

Interview Findings

Despite our efforts to prioritize the inclusion of stuttering advocates with prior experience in stuttering data collection in this study, we found that only a few interviewees had the time or capacity to meaningfully engage with data governance frameworks or the data justice movement prior to our study. Nevertheless, there was a shared recognition of the need for representative stuttered speech data to support the development of stuttering-friendly speech AI, as well as enthusiasm for stutterer-led, community-centered approaches to data collection and governance. Alongside the enthusiasm, interviewees offered important considerations and constructive insights on how to collect, share, and safeguard community data responsibly.

Motivating Data Contribution: Transparency and Community Benefits

We identified a shared belief among advocates that people who stutter are typically generous in contributing their voices to community-benefiting initiatives, when the benefits and goals are clearly communicated to them. Alex emphasized that PWS are more motivated to contribute when their participation is clearly linked to meaningful contributions to their own community.

> *The biggest thing is clearly communicating that participants have a real opportunity to meaningfully impact the development of the technology and play a significant role in how it's created. . . . If it's just framed as, "We're collecting speech samples for X research project," it can feel unclear or unmotivating. But if it's presented as, "You have the chance to be one of 50 people helping to create accessible AI for people who stutter," that's exciting.*

One example of articulating the meaningful role contributors would play is the *Library of Disfluent Voices*, an open-source collection of voice recordings voluntarily contributed by PWS to celebrate stuttering and speech diversity. Led by a stuttering advocacy and support organization, this initiative has attracted contributors from various ages, genders, and racial and ethnic backgrounds, many of whom use their real names. Despite the potential privacy risks of open-sourcing speech data, Alex expressed,

> *The rewards likely outweigh the risks... People knew when they were recording that they were contributing to a public project, that their voices would be used to create music, and that their recordings would be available online. We maintained transparency and clear communication around the process.*

This sentiment aligns with findings from previous research (Li and Wu 2024), highlighting the community's constructive agency and the need to restructuring the AI data ecosystem – from merely consuming community data to collaboratively building with the community.

Building Ethical Consent: Bridging Community Goals and Resource Constraints

Informed consent emerged as a fundamental component of ethical practice in our interviews. Natalie shared consent forms she developed for an established clinical speech dataset platform which outlines critical information clearly, including the study's purpose, eligibility criteria, detailed procedures, potential benefits, anticipated risks and discomforts, confidentiality guarantees, compensation, the voluntary nature of participation, participants' rights to withdraw, and essential contact information for inquiries. She emphasized that adopting a similar consent process will ensure participants are thoroughly informed and empowered to make autonomous decisions regarding their data.

Beyond the basic consent, several of our advocates emphasized the importance of tiered consent for participant-centered data governance. Rather than a one-size-fits-all agreement, participants should be granted fine-grained control over what types of data (e.g., audio, video) they contribute and how the data is stored, shared, or made public. As Nancy suggested,

> If you're hoping to contribute some stuttered voices to [a] database that's open access, I wonder if there are different tiers that a participant can agree to, like if they agree to both audio and video being put into the database, or if they just want audio ... giving people some choices about what they're okay with.

While Natalie and Nancy were able to leverage established ethical frameworks and institutional support such as the IRB to implement effective, informed consent for their data collection participants, grassroots, community-led initiatives often lack the resources and the know-hows needed to design an effective consent process that meets their values and goals. Alex, for instance, shared his experience developing the release form for the *Library of Disfluent Voices*. Due to limited time and resources, the initial version was assembled using generative AI tools with small adjustments to meet pressing project needs, *"We had to get some stock language up because we were running out of time and needed participants to sign. Most of it came from ChatGPT, which we then adjusted."* As a result of this ad-hoc approach, the release form of *Disfluent Voices* uses generic yet comprehensive consent language that allows broad future uses of the contributed data, and some of which might not align with the community's expectations and goals. As Alex later noticed,

> When we filled out the form, I didn't think that at some point, a tech company might want to use these recordings as well for a different project.

While Alex planned to seek legal consultation to enhance the consent form for *Disfluent Voices*, he acknowledged the resource limitations that hinder this process.

This practical challenge in designing a community-centered consent process not only undermines the community's values and goals, but also introduces additional liabilities and risks for individuals in the community. For instance, Monica used a release form adapted from a filmmaker friend to record video and audio data of people who stutter for her advocacy podcast and documentary film. However, the legal liability currently falls on her personally, and she recognizes the need to form a limited liability company (LLC) for legal protection in the future – once she has the necessary resources and capacity.

These experiences underscores a critical need for accessible legal guidance, consent templates, and operational support tailored for grassroots and low-resourced communities that enable them to both participate in and design respectful, ethical data consent process.

Safeguarding Community Data: Navigating Open Access and Heighten Privacy Risks

The community's willingness to engage in community-centered data initiatives does not imply indifference to the associated privacy and data security risks. On the contrary, advocates emphasized the importance of safeguarding, especially when dealing with sensitive, identity-linked speech data. For example, advocates espoused essential practices such as data de-identification, straightforward mechanisms for participants to request data deletion, and ongoing, transparent monitoring on how data is utilized. Nancy emphasized the inherent complexity involved in managing these competing needs: *"It is a challenge balancing open science with protecting participants' confidentiality."*

Additionally, advocates underscored the need for controlled access to sensitive datasets. Four had experience with publishing speech data through access-by-request mechanisms hosted on project websites. Natalie elaborated on the benefits of password-protected mechanism for disfluent speech datasets, suggesting that even partial restrictions on data access can reassure potential contributors:

> We still have password-protected data, and while it might technically be open access, it's not that you simply ask for a user account and then everything is open to you...that may help you get the data collected, because people have less to worry about if data is password protected.

While this approach prioritize data protection, these advocates also believed that stuttered speech data holds greater value when shared with researchers and companies actively working on stuttering-friendly technologies, and thus should be made accessible in ways that both support innovation and respect contributors' rights and dignity.

To address this, Natalie recommended sharing community datasets through established platforms with controlled access that allows contributors to retain a degree of control through community approved terms of use for third-party data users. This approach consists of a request and approval process to access the data, requirement to cite and attribute of dataset creators, and explicit prohibition of certain use cases that fall outside the consented scope. As she explained:

> The data cannot be used without some sort of permission... it will ask that people who use it have to cite it... and then uses other than, for instance, teaching and software development are expressly prohibited.

Natalie also shared an example where a non-community member requested access to the speech data for an art instal-

lation, which Natalie denied. This decision was driven by the incongruence between the use case and the purpose for which participants consented to share their data.

> *We had a request that we turned down where somebody said, I want to make an art installation... and it's like no, no. That is not why people gave us these recordings.*

The access and privacy considerations around speech data can be further complicated by geo-political tensions. Ray and Teresa shared the significant challenges they faced in sharing the Chinese stuttered speech data they collected with researchers outside China, especially in navigating regulations around cross-border data transfer and regional personal data laws. It took them nearly a year of time and significant legal resources to address these issues.

Beyond community-controlled bulk data access and sharing, our interviewees also shared their existing practices and recommendations for protecting the privacy of individual data contributors.

Natalie emphasized the importance of offering simple and accessible mechanisms for data deletion, given the sensitivity of speech and video data of PWS. While data contributors may initially consent to share their recordings, their comfort levels and expectations may evolve over time—especially as they see how their data is used or consider potential future contexts of use. Natalie shared the example of receiving a request of data removal from a data contributor years later, and how she willingly complied to such requests. This example underscores the need for infrastructure that supports data contributors' decision making about their data long after data have been collected and shared.

Ray, Terasa, and Natalie also described their approach to de-identify speakers in stuttered speech data by manually redacting names and other personal identifiable information from recordings and transcripts. While this process is highly labor-intensive and time-consuming, the effectiveness of such measures remains uncertain—especially given the uniquely identifiable nature of stuttering and voice data.

To summarize, while the advocates recognize the community's needs for comprehensive privacy protection and fine-grained, long-term control of their data, they currently lack the tools and infrastructures to effectively meet these needs while supporting open access and broader sharing of the community data.

Transparency Beyond Collection: Keeping Participants Informed

Besides during the initial data collection, the importance of transparency in data governance was also underscored by Eric's prior experience participating in a speech recording study. While he was willing to contribute his data to improve speech technologies, he was never informed about the outcomes or the specific way that his data was used. This absence of transparency and communication left him uncertain about the broader impact and benefits of his involvement. Similar experience and sentiment was shared by Monica. Nancy addressed this concern, stressing the ethical respon-

sibility of researchers to ensure participants maintain autonomy and clarity regarding the use of their data:

> *We need to tread very carefully and ensure that participants have full autonomy over what they are comfortable sharing and how. It's not just about sharing for the sake of a current project, but also about how their data is stored and shared beyond that.*

Nancy's perspective highlights the need for transparent and sustained engagement with participants beyond the initial data collection. Such engagement can include ongoing updates about data usage, storage practices, and any secondary applications. In particular, Jane recommended establishing persistent and proactive communication channels with participants and the broader community. For example, she suggested providing regular and accessible updates about the project's status, milestones, and outcomes via a website: *"If you have a website, people can check the progress of the project."* Such transparency would empower participants, fostering a deeper sense of ownership and involvement.

Survey Results

Attitudes Towards Stuttering

To understand the societal contexts and constraints around stuttering and data sharing, we first assess the participants' general attitudes towards stuttering. Aligned with previous findings (Ip et al. 2012; Ma et al. 2023), Chinese respondents reported higher self- and social stigma, which may disincentivize data contribution.

Negative thoughts toward stuttering. Participants were asked "*how many negative thoughts or feelings do you generally have about stuttering? (1 = None, 5 = Many)*". On average, Chinese participants reported significantly more negative thoughts toward stuttering ($Mean = 3.60$) compared to U.S. participants ($Mean = 3.03$). As shown in 1, 71.4% of Chinese participants (60 out of 84) reported substantial negative thoughts (i.e. "Quite a few" or "Many") towards stuttering, compared to 40.9% in the US participants, indicating greater internalized stigma and lower public acceptance in Chinese sample.
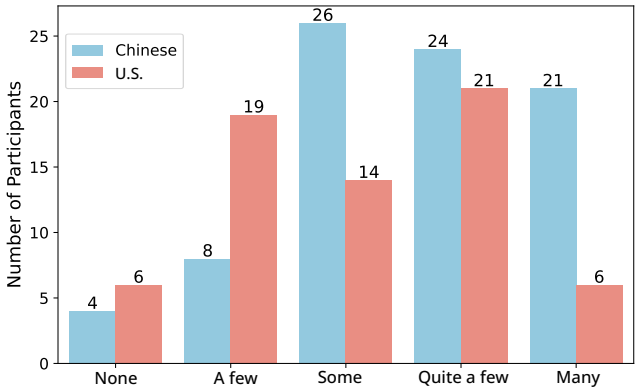

Figure 1: Distribution of negative thoughts towards stuttering.

Frequency of avoiding stuttering. Participants rated how frequently they avoid speaking situations due to stuttering on a 5-point Likert scale (1 = never, 5 = always). Fig. 2 shows 61.9% of Chinese participants (52/84) and 53.0% of U.S. participants (35/66) reported avoiding often or always. The average avoidance score was also higher for Chinese participants ($Mean = 3.76$) than for U.S. participants ($Mean = 3.38$). This suggests that Chinese participants are more likely to mask their stuttering and identity as PWS.
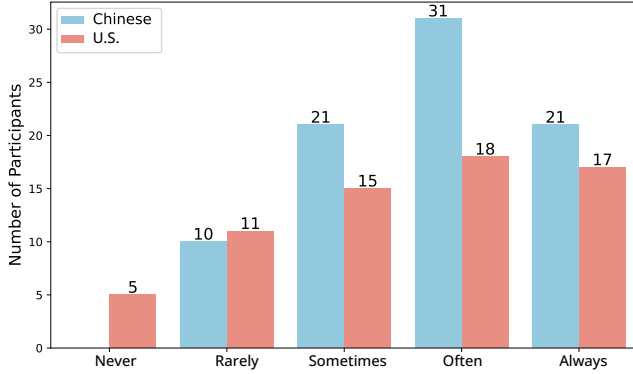
Figure 2: Distribution of frequency of avoiding stuttering.

Considerations for Data Contribution

Despite different levels of societal pressure around stuttering, participants from both countries share similar concerns and considerations when contributing their speech data.

Participants were asked to rate the importance of four factors that could influence their willingness to share data: being informed on who has access to their data, receiving regular updates on research or project outcomes, receiving compensation, and anonymization of data. As shown in Fig. 3, while all four factors were rated as important by participants from both countries, statistically significant higher ratings were observed in Chinese participants in three factors: *receiving regular updates* ($U = 1625.50, p < 0.0001$), *receiving compensation for data contribution* ($U = 1980.50, p = 0.0030$), and *data anonymization* ($U = 2006.00, p = 0.0021$). No significant difference was found for the importance of being informed about who has access to the data ($U = 2852.50, p = 0.6197$), which suggests that transparency around data access is equally valued by both groups. Also, *receiving compensation* was rated as the least important factor by participants in both countries, confirming the finding on intrinsic over extrinsic motivation reported in previous research (Li and Wu 2024).

Data Sharing Preferences and Concerns

To better inform the decision on data share and use cases, we asked the participants to rate on their preferences to share data with different entities and for different purposes, and to share their biggest concerns around data sharing.

Preferences for Data Users. Participants rated their willingness to share data with five types of institutions: *univer-*

sities and academic researchers, *nonprofit organizations focused on stuttering*, *government agencies*, *for-profit companies*, and *media agencies*. These options were drawn from previous research on data share and stewardship (Tseng et al. 2024; Berke et al. 2024). Both Chinese and U.S. groups showed a clear preference for sharing data with universities and stuttering-focused nonprofit organizations, while expressing greater hesitation to share data with the media and for-profit companies. Despite earlier findings on Chinese citizens' stronger acceptance of government use of biometric data (Kostka, Steinacker, and Meckel 2021), we find data use by government agencies occupies a middle ground for both groups.

Preferences for Data Usage Goals.

Data Protection and Governance Measures

Participants were asked to rate their level of acceptance regarding the use of their speech data for various purposes, including *improving technology systems*, *supporting training for Speech-Language Pathologists*, *social advocacy to raise public awareness about stuttering*, *academic research*, and *commercial product development*. These options were derived from our interview study with community advocates. Overall, participants across both groups were supportive of data sharing for public benefit, including technology improvement, training speech language pathologists, and advocacy. We observed significant differences emerged for academic research ($U = 2265.50, p = 0.0190$) and especially for commercial product development ($U = 1637.00, p < 0.0001$), where Chinese participants demonstrated greater acceptance than their U.S. counterparts, as shown in Fig 4.

Concerns About the Risks of Data Sharing. Participants were asked to rate their level of concern regarding three potential risks of data sharing: (1) data being used to develop discriminatory technologies against PWS, (2) leakage of private and sensitive information, and (3) data being used to create harmful AI-generated content. These potential risks were drawn from discussions with advocates in our interview study. As shown in Fig 5, participants from China and the US shared moderate to high levels of concern across all three categories. Mean concern scores were slightly higher for the Chinese community – for example, concern about discriminatory technologies averaged 3.83 among Chinese participants and 3.44 among US participants.

Participants were asked to evaluate the importance of four data protection and governance measures: (1) pre-screening data users' qualifications and motivations, (2) receiving regular updates on how the data is used, (3) open communication channels for community feedback and questions, and (4) easy options for deleting personal data. Overall, both Chinese and U.S. participants rated these measures as important, with mean scores clustering between 4.0 and 4.4 on a 5-point scale. Chinese participants rated regular updates on data use significantly more important than U.S. participants ($U = 2099.00, p = 0.0098$), which suggests a stronger expectation for ongoing transparency and accountability. No statistically significant differences were observed for the other three items.
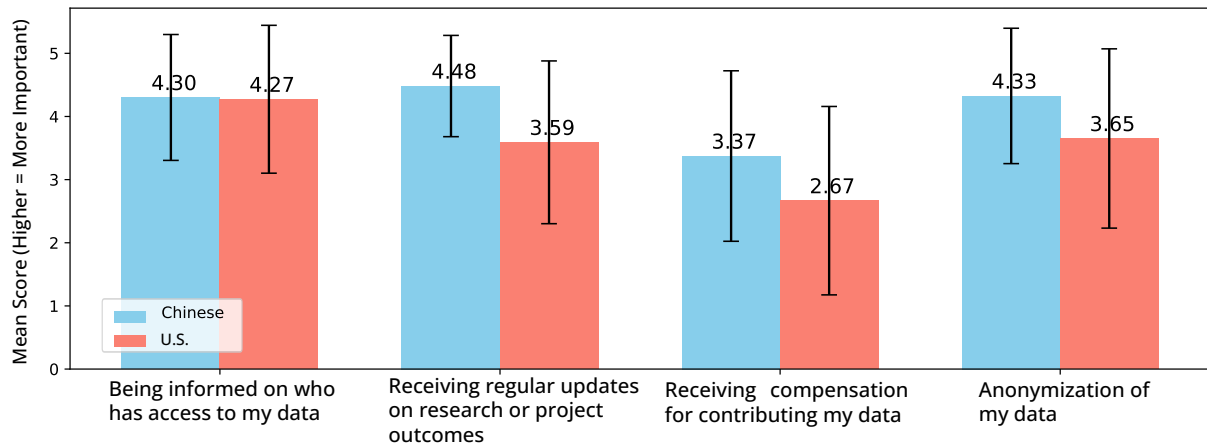
Figure 3: Mean importance scores for factors influencing willingness to contribute data by country.
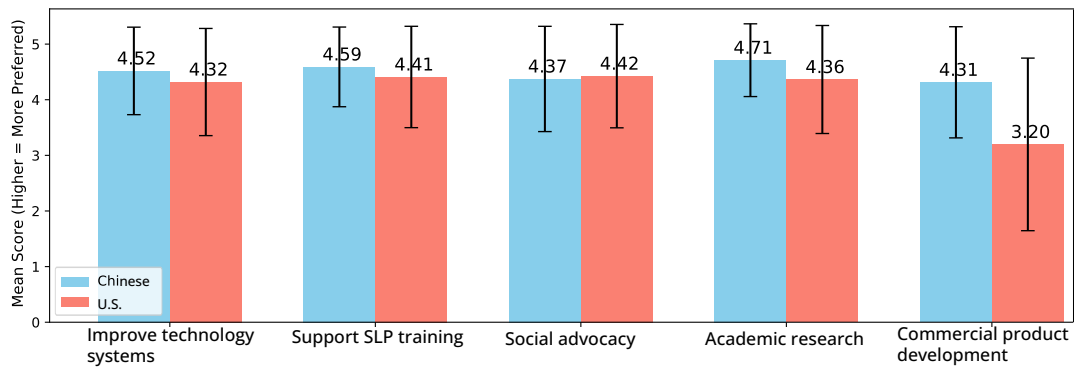

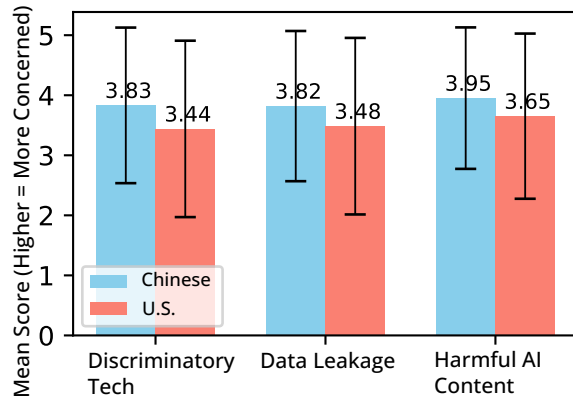Figure 4: Mean acceptance scores for different data-sharing purposes by country.

Figure 5: Concerns about the risks of data sharing.

Discussion

Community-Centered Disability Data Governance

This study extends prior literature by exploring how trusted relationships within the stuttering community can translate into governance preferences for AI data. Prior work showed how direct involvement of PWS in the design and develop-

ment process not only empowered contributors, but also fostered sustained trust in the organizational stewards(Li and Wu 2024; Li, Wu, and Leshed 2024; Li et al. 2025). Our findings suggest that this trust carries over into preferences around data governance: participants were generally more comfortable sharing data with universities and nonprofit organizations than with commercial or media entities, and emphasized priorities such as transparency, consent, and contributor agency.

Our study aligns with longstanding critiques that people with disabilities are often excluded from meaningful roles in shaping how their data is collected, governed, and used, despite being central to the functioning and evaluation of AI systems (Zuboff 2023; Tuck and Yang 2014; Park et al. 2021). Both our interviews and surveys revealed strong community interest in being involved throughout the full data lifecycle. Participants emphasized a desire for control mechanisms such as flexible consent, contributor review of reuse requests, and regular updates on data applications.

These preferences resonate with recent calls for "disability-first" and participatory data stewardship models (Theodorou et al. 2021; Whittaker et al. 2019), and align with the broader shift in data justice literature toward governance as an ongoing relationship rather than a static act
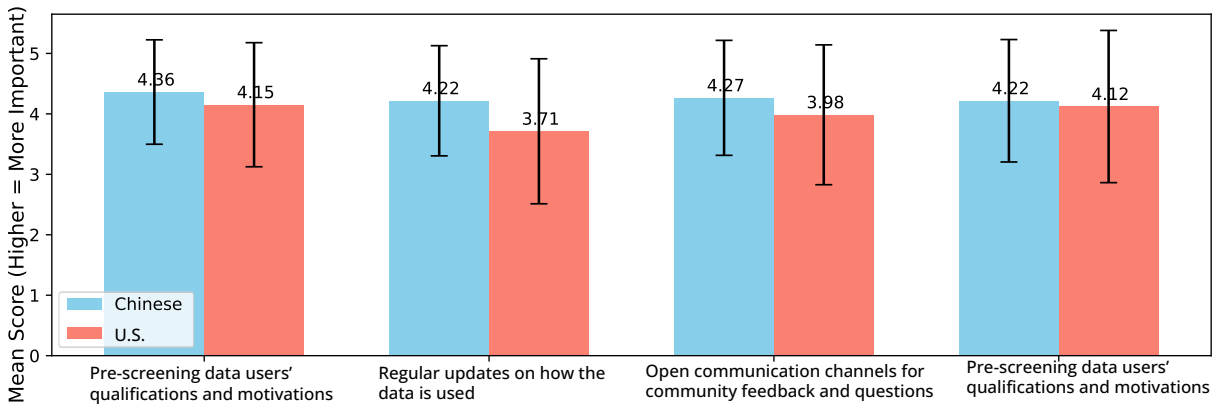
Figure 6: Mean importance ratings of data protection measures by country. Chinese participants rated "regular updates on how the data is used" significantly higher than the U.S. participants ($p = 0.0098$).

(Tseng et al. 2024; Tran et al. 2022). Our findings also echo Erete et al.'s recommendation to "set the table" early in community-tech collaborations by formalizing shared values, transparency commitments, and IP arrangements (Erete et al. 2025). Such grounding ensures that data partnerships are built on trust and reciprocity, not just access. As a trusted steward, we view our role not simply as protecting data privacy, but as facilitating a governance structure that centers the community's agency and evolving priorities.

Variation in Social Stigma, Convergence in Governance Values

While our interviews and surveys highlight socio-geographical differences in how stuttering is experienced—especially the heightened stigma and avoidance behaviors reported among Chinese participants—our analysis nonetheless showed remarkable similarity in data governance preferences. In both China and the US, participants favored data uses that directly benefit the stuttering community, such as ASR technology improvement, advocacy, and speech therapy. This similarity is notable given the substantial differences in actual risk: technological surveillance is more pervasive in China, and legal remedies for data misuse are more limited, whereas in the U.S., protections are fragmented but civil society and advocacy groups play a more active oversight role. A possible explanation is that perceptions of risk may stem less from objective harm and more from shared experiences of marginalization. In both contexts, PWS face similar barriers in speech technologies, stigma, and limited agency over their data, which may outweigh differences in surveillance intensity.

Furthermore, there was a shared concern on the use of stuttered speech data by commercial or media entities, especially without community oversight or control. This reflects a growing awareness of the risks posed by extractive data practices, including misuse, misrepresentation, and loss of control (Zuboff 2023; Benjamin 2018).

Limitations and Future Work

As we center the stuttering community's perspectives in the collection and governance of their own data, we recognize that many participants, even stuttering advocates, lacked prior experience with data governance frameworks or technical expertise in AI. As a result, participants often expressed their values and goals in broad terms rather than through specific legal or technical requirements, echoing prior findings that historically marginalized communities lack literacy or opportunity to engage in decisions around data use and infrastructure (Wong 2023; Lin et al. 2024; Erete et al. 2025). This highlights the need for future work to translate community values into actionable, enforceable governance mechanisms, such as consent structures, access protocols, and terms of use agreement—the unglamorous but essential details for implementing inclusive, community-driven governance.

Second, while we strove to make our interview and survey questions accessible and grounded in real-world examples, we acknowledge that the design of governance systems involves complex trade-offs that are difficult to capture through short-term engagements Our engagement methods were limited by varying levels of digital literacy and prior experience with the abstract concept of data governance. Future work should explore participatory design methods that scaffold literacy around data use and sharing such as through scenario-based workshops or speculative design.

Conclusion

This paper examines the ethical governance of stuttered speech data through a disability-centered, community-led lens, drawing on interviews and surveys with people who stutter in the US and China. Our findings highlight that despite socio-geographical differences, participants share common governance values—emphasizing transparency, agency, and protection from harm. Our study offers important insights into community preferences and priorities around the governance of stuttered speech data, calling for a shift from extractive practices to participatory stewardship.

Positionality Statement

We acknowledge that our personal backgrounds and identities shape how we engage with communities and interpret our findings. The first, second, fourth and last authors are Mandarin-speaking researchers and had experience living and studying in China. Our research team includes both PWS and non-stuttering allies. The last author identifies as a person who stutters and maintains both personal and professional connections within the stuttering communities in the United States and China, which enabled the research team to engage participants who are deeply involved in stuttering advocacy, research, and grassroots.

References

Ardila, R.; Branson, M.; Davis, K.; Kohler, M.; Meyer, J.; Henretty, M.; Morais, R.; Saunders, L.; Tyers, F.; and Weber, G. 2020. Common Voice: A Massively-Multilingual Speech Corpus. In Calzolari, N.; Béchet, F.; Blache, P.; Choukri, K.; Cieri, C.; Declerck, T.; Goggi, S.; Isahara, H.; Maegaard, B.; Mariani, J.; Mazo, H.; Moreno, A.; Odijk, J.; and Piperidis, S., eds., *Proc. Twelfth Lang. Resour. Eval. Conf.*, 4218–4222. Marseille, France: European Language Resources Association. ISBN 979-10-95546-34-4.

Bellman, S.; Johnson, E. J.; Kobrin, S. J.; and Lohse, G. L. 2004. International differences in information privacy concerns: A global survey of consumers. *Inf. Soc.*, 20(5): 313–324.

Bender, E. M.; Gebru, T.; McMillan-Major, A.; and Shmitchell, S. 2021. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? . In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, 610–623. New York, NY, USA: Association for Computing Machinery. ISBN 9781450383097.

Benjamin, R. 2018. Science, Politics, and Group-Making in the Postcolony. *Reconsidering Race: Social Science Perspectives on Racial Categories in the Age of Genomics*, 173.

Benjamin, R. 2023. Race after technology. In *Social Theory Re-Wired*, 405–415. Routledge.

Berke, A.; Mahari, R.; Pentland, A.; Larson, K.; and Calacci, D. 2024. Insights from an Experiment Crowdsourcing Data from Thousands of US Amazon Users: The importance of transparency, money, and data use. *Proc. ACM Hum.-Comput. Interact.*, 8(CSCW2).

Bernstein Ratner, N.; and MacWhinney, B. 2018. Fluency Bank: A New Resource for Fluency Research and Practice. *Journal of Fluency Disorders*, 56: 69–80.

Bigham, J. P.; Jayant, C.; Ji, H.; Little, G.; Miller, A.; Miller, R. C.; Miller, R.; Tatarowicz, A.; White, B.; White, S.; and Yeh, T. 2010. VizWiz: nearly real-time answers to visual questions. In *Proceedings of the 23nd annual ACM symposium on User interface software and technology*. New York, NY, USA: ACM.

Bloodstein, O.; Ratner, N. B.; and Brundage, S. B. 2021. *A Handbook on Stuttering*. San Diego, CA: Plural Publishing, Inc, seventh edition edition. ISBN 978-1-63550-318-0.

Buyl, M.; Cociancig, C.; Frattone, C.; and Roekens, N. 2022. Tackling algorithmic disability discrimination in the hiring process: An ethical, legal and technical analysis. In *2022 ACM Conference on Fairness, Accountability, and Transparency*. New York, NY, USA: ACM.

Charmaz, K. 2014. *Constructing grounded theory*. SAGE publications Ltd.

Coalson, G. A.; Crawford, A.; Treleaven, S. B.; Byrd, C. T.; Davis, L.; Dang, L.; Edgerly, J.; and Turk, A. 2022. Microaggression and the Adult Stuttering Experience. *Journal of Communication Disorders*, 95: 106180.

Constantino, C.; Campbell, P.; and Simpson, S. 2022. Stuttering and the Social Model. *Journal of Communication Disorders*, 96: 106200.

Costanza-Chock, S. 2020. *Design justice: Community-led practices to build the worlds we need*. The MIT Press.

Dencik, L. 2022. Data and Social Justice. *Data Justice. Los Angeles: Sage. S*, 123–137.

Erete, S.; Corbett, E.; Smith-Walker, N.; Cunningham, J. L.; Gatz, E.; Park, T. M.; Perry, T.; Wilcox, L.; and Denton, R. 2025. Towards equitable community-industry collaborations: Understanding the experiences of nonprofits' collaborations with tech companies. *Proceedings of the ACM on Human-Computer Interaction*, 9: 1–31.

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

Foley, R. W.; Sylvain, O.; and Foster, S. 2022. Innovation and equality: an approach to constructing a community governed network commons. *Journal of Responsible Innovation*, 9(1): 49–73.

Ggottermeier Skushalnagar Raja, L. 2016. User Evaluation Automatic Speech Recognition Systems Deaf-Hearing Interactions School Work. *Audiology Today*, 28: 20–34.

Glazko, K.; Mohammed, Y.; Kosa, B.; Potluri, V.; and Mankoff, J. 2024. Identifying and improving disability bias in GPT-based resume screening. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*, 687–700. New York, NY, USA: ACM.

Green, J. R.; MacDonald, R. L.; Jiang, P.-P.; Cattiau, J.; Heywood, R.; Cave, R.; Seaver, K.; Ladewig, M. A.; Tobin, J.; Brenner, M. P.; Nelson, P. C.; and Tomanek, K. 2021. Automatic Speech Recognition of Disordered Speech: Personalized Models Outperforming Human Listeners on Short Phrases. In *Interspeech 2021*, 4778–4782. ISCA.

Guo, A.; Kamar, E.; Vaughan, J. W.; Wallach, H.; and Morris, M. R. 2019. Toward fairness in AI for people with disabilities: A research roadmap. *arXiv [cs.CY]*.

Gurari, D.; Li, Q.; Stangl, A. J.; Guo, A.; Lin, C.; Grauman, K.; Luo, J.; and Bigham, J. P. 2018. VizWiz grand challenge: Answering visual questions from blind people. *arXiv [cs.CV]*.

Hsieh, J.; Zhang, A.; Kim, S.; Rao, V. N.; Dalal, S.; Mateescu, A.; Grohmann, R. D. N.; Eslami, M.; and Zhu, H. 2024. Worker Data Collectives as a means to Improve Accountability, Combat Surveillance and Reduce Inequalities. In *Companion Publication of the 2024 Conference on Computer-Supported Cooperative Work and Social Computing*, 697–700.

Ip, M. L.; St. Louis, K. O.; Myers, F. L.; and Xue, S. A. 2012. Stuttering Attitudes in Hong Kong and Adjacent Mainland China. *International Journal of Speech-Language Pathology*, 14(6): 543–556.

Jamieson, S. 2004. Likert scales: How to (ab) use them? *Medical education*, 38(12): 1217–1218.

Kamikubo, R.; Lee, K.; and Kacorri, H. 2023. Contributing to accessibility datasets: Reflections on sharing study data by blind people. *Proc. SIGCHI Conf. Hum. Factor. Comput. Syst.*, 2023: 827.

Kamikubo, R.; Zamiri Zeraati, F.; Lee, K.; and Kacorri, H. 2024. AccessShare: Co-designing data access and sharing with blind people. In *The 26th International ACM SIGACCESS Conference on Computers and Accessibility*, volume 4, 1–16. New York, NY, USA: ACM.

Kiger, G. 1992. Disability simulations: Logical, methodological and ethical issues. *Disabil. Handicap Soc.*, 7(1): 71–78.

Kostka, G.; Steinacker, L.; and Meckel, M. 2021. Between security and convenience: Facial recognition technology in the eyes of citizens in China, Germany, the United Kingdom, and the United States. *Public Understanding of Science*, 30(6): 671–690. PMID: 33769157.

Kourkounakis, T.; Hajavi, A.; and Etemad, A. 2021. FluentNet: End-to-End Detection of Stuttered Speech Disfluencies With Deep Learning. *IEEE/ACM Trans. Audio, Speech and Lang. Proc.*, 29: 2986–2999.

Kuzio, J.; Ahmadi, M.; Kim, K.-C.; Migaud, M. R.; Wang, Y.-F.; and Bullock, J. 2022. Building better global data governance. *Data 38; Policy*, 4: e25.

Lea, C.; Huang, Z.; Narain, J.; Tooley, L.; Yee, D.; Tran, D. T.; Georgiou, P.; Bigham, J. P.; and Findlater, L. 2023. From User Perceptions to Technical Improvement: Enabling People Who Stutter to Better Use Speech Recognition. In *Proc. 2023 CHI Conf. Hum. Factors Comput. Syst.*, CHI '23, 1–16. New York, NY, USA: Association for Computing Machinery. ISBN 978-1-4503-9421-5.

Lea, C.; Mitra, V.; Joshi, A.; Kajarekar, S.; and Bigham, J. P. 2021. SEP-28k: A Dataset for Stuttering Event Detection from Podcasts with People Who Stutter. In *ICASSP 2021 - 2021 IEEE Int. Conf. Acoust. Speech Signal Process. ICASSP*, 6798–6802.

Leslie, D.; Katell, M.; Aitken, M.; Singh, J.; Briggs, M.; Powell, R.; Rincon, C.; Perini, A.; and Jayadeva, S. 2022. Data justice in practice: A guide for impacted communities. *SSRN Electron. J.*

Li, J.; Li, Q.; Gong, R.; Wang, L.; and Wu, S. 2025. Our Collective Voices: The Social and Technical Values of a Grassroots Chinese Stuttered Speech Dataset. *The 2025 ACM Conference on Fairness, Accountability, and Transparency (FAccT '25), June 23-26, 2025, Athens, Greece*, 1(1).

Li, J.; Wu, S.; and Leshed, G. 2024. Re-envisioning Remote Meetings: Co-designing Inclusive and Empowering Video-conferencing with People Who Stutter. In *Proceedings of the 2024 ACM Designing Interactive Systems Conference*, DIS '24, 1926–1941. New York, NY, USA: Association for Computing Machinery. ISBN 9798400705830.

Li, Q.; and Wu, S. 2024. "I want to publicize my stutter": Community-led collection and curation of Chinese stuttered speech data. *Proc. ACM Hum. Comput. Interact.*, 8(CSCW2): 1–27.

Lin, H.; Karusala, N.; Okolo, C. T.; D'Ignazio, C.; and Gajos, K. Z. 2024. "Come to us first": Centering Community Organizations in Artificial Intelligence for Social Good Partnerships. *Proceedings of the ACM on Human-Computer Interaction*, 8(CSCW2): 1–28.

Ma, Y.; Oxley, J. D.; Yaruss, J. S.; and Tetnowski, J. A. 2023. Stuttering experience of people in China: A cross-cultural perspective. *Journal of Fluency Disorders*, 77: 105994.

MacDonald, R. L.; Jiang, P.-P.; Cattiau, J.; Heywood, R.; Cave, R.; Seaver, K.; Ladewig, M. A.; Tobin, J.; Brenner, M. P.; Nelson, P. C.; Green, J. R.; and Tomanek, K. 2021. Disordered Speech Data Collection: Lessons Learned at 1 Million Utterances from Project Euphonia. In *Interspeech 2021*, 4833–4837. ISCA.

Paris, D.; and Winn, M. T. 2013. *Humanizing research: Decolonizing qualitative inquiry with youth and communities*. Sage Publications.

Park, J. S.; Bragg, D.; Kamar, E.; and Morris, M. R. 2021. Designing an online infrastructure for collecting AI data from people with disabilities. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. New York, NY, USA: ACM.

Radford, A.; Kim, J. W.; Xu, T.; Brockman, G.; McLeavey, C.; and Sutskever, I. 2023. Robust speech recognition via large-scale weak supervision. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org.

Saldaña, J. 2021. *The coding manual for qualitative researchers*. SAGE publications Ltd.

Sharma, T.; Stangl, A.; Zhang, L.; Tseng, Y.-Y.; Xu, I.; Findlater, L.; Gurari, D.; and Wang, Y. 2023. Disability-first design and creation of A dataset showing private visual information collected with people who are blind. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, volume 79, 1–15. New York, NY, USA: ACM.

Silverman, A. M.; Gwinn, J. D.; and Van Boven, L. 2015. Stumbling in their shoes: Disability simulations reduce judged capabilities of disabled people. *Soc. Psychol. Personal. Sci.*, 6(4): 464–471.

Smith, H. J.; Dinev, T.; and Xu, H. 2011. Information privacy research: An interdisciplinary review. *MIS Q.*, 35: 989–1015.

Taylor, L. 2017. What is data justice? The case for connecting digital rights and freedoms globally. *Big Data & Society*, 4(2): 2053951717736335.

The National People's Congress of PRC . 2021. Personal Information Protection Law of the People's Republic of China. http://en.npc.gov.cn.cdurl.cn/2021-12/29/c_694559.htm. Accessed: 2025-08-07.

Theodorou, L.; Massiceti, D.; Zintgraf, L.; Stumpf, S.; Morrison, C.; Cutrell, E.; Harris, M. T.; and Hofmann, K. 2021. Disability-first dataset creation: Lessons from constructing a dataset for teachable object recognition with blind and low vision data collectors. In *Proceedings of the 23rd International ACM SIGACCESS Conference on Computers and Accessibility*. New York, NY, USA: ACM.

Tichenor, S. E.; and Yaruss, J. S. 2021. Variability of Stuttering: Behavior and Impact. *Am J Speech Lang Pathol*, 30(1): 75–88.

Tobin, J.; and Tomanek, K. 2022. Personalized Automatic Speech Recognition Trained on Small Disordered Speech Datasets. In *ICASSP 2022 - 2022 IEEE Int. Conf. Acoust. Speech Signal Process. ICASSP*, 6637–6641.

Tran, A.-T.; Boone, A.; Le Dantec, C. A.; and DiSalvo, C. 2022. Careful Data Tinkering. *Proceedings of the ACM on Human-Computer Interaction*, 6(CSCW2): 1–29.

Tseng, E.; Bellini, R.; Lee, Y.-Y.; Ramjit, A.; Ristenpart, T.; and Dell, N. 2024. Data Stewardship in Clinical Computer Security: Balancing Benefit and Burden in Participatory Systems. *Proceedings of the ACM on Human-Computer Interaction*, 8(CSCW1): 1–29.

Tuck, E.; and Yang, K. W. 2014. R-words: Refusing research. *Humanizing research: Decolonizing qualitative inquiry with youth and communities*, 223(2014): 248.

US Privacy and Civil Liberties Oversight Board. 2025. PCLOB. https://www.pclob.gov/. Accessed: 2025-08-07.

Wenzel, K.; Devireddy, N.; Davison, C.; and Kaufman, G. 2023. Can Voice Assistants Be Microaggressors? Cross-Race Psychological Responses to Failures of Automatic Speech Recognition. In *Proc. 2023 CHI Conf. Hum. Factors Comput. Syst.*, CHI '23, 1–14. New York, NY, USA: Association for Computing Machinery. ISBN 978-1-4503-9421-5.

Whittaker, M.; Alper, M.; Bennett, C. L.; Hendren, S.; Kaziunas, L.; Mills, M.; Morris, M. R.; Rankin, J.; Rogers, E.; Salas, M.; and Others. 2019. Disability, bias, and AI. *AI Now Institute*, 8: 11.

Wong, J. 2023. Data practices and data stewardship. *Interactions*, 30(3): 60–63.

Wu, S.; Reynolds, L.; Li, X.; and Guzmán, F. 2019. Design and evaluation of a social media writing support tool for people with dyslexia. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. New York, NY, USA: ACM.

Zuboff, S. 2023. The age of surveillance capitalism. In *Social theory re-wired*, 203–213. Routledge.