AAAI / ACM conference on
ARTIFICIAL INTELLIGENCE, ETHICS, AND SOCIETY

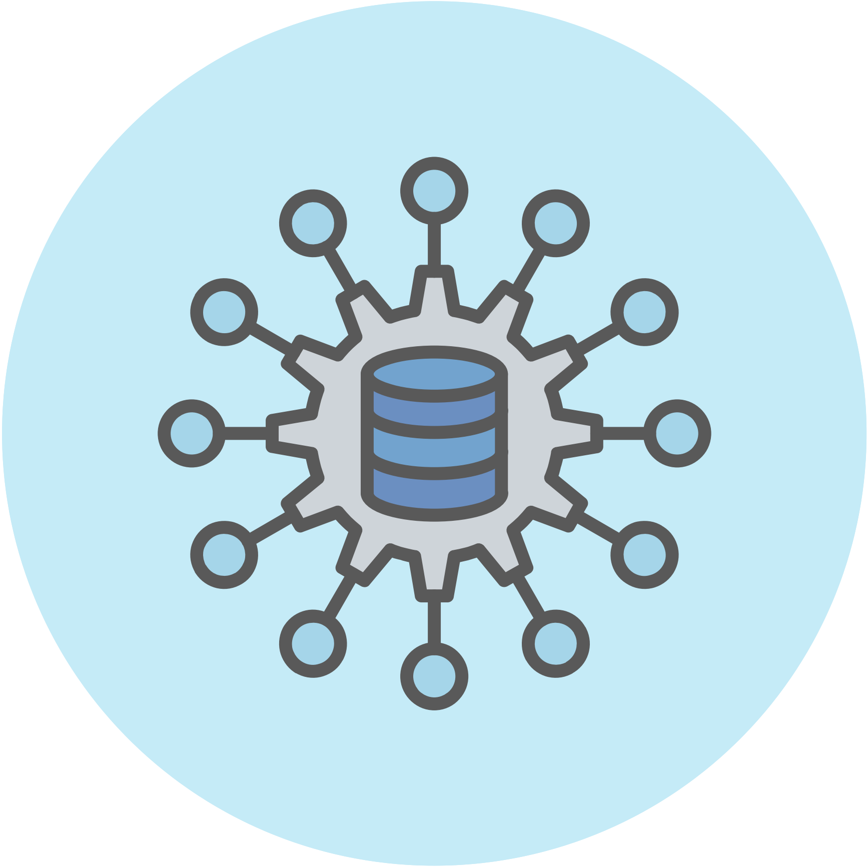
Govern With, Not For:

Understanding the Stuttering Community's Preferences and Goals for Speech AI Data Governance in the US and China

Jingjin Li, AImpower.org, USA
Peiyao Liu, University of California, Santa Cruz, USA
Rebecca Lietz, University of California, Santa Cruz, USA
Ningjing Tang, Carnegie Mellon University, USA
Norman Makoto Su, University of California, Santa Cruz, USA
Shaomei Wu, AImpower.org, USA

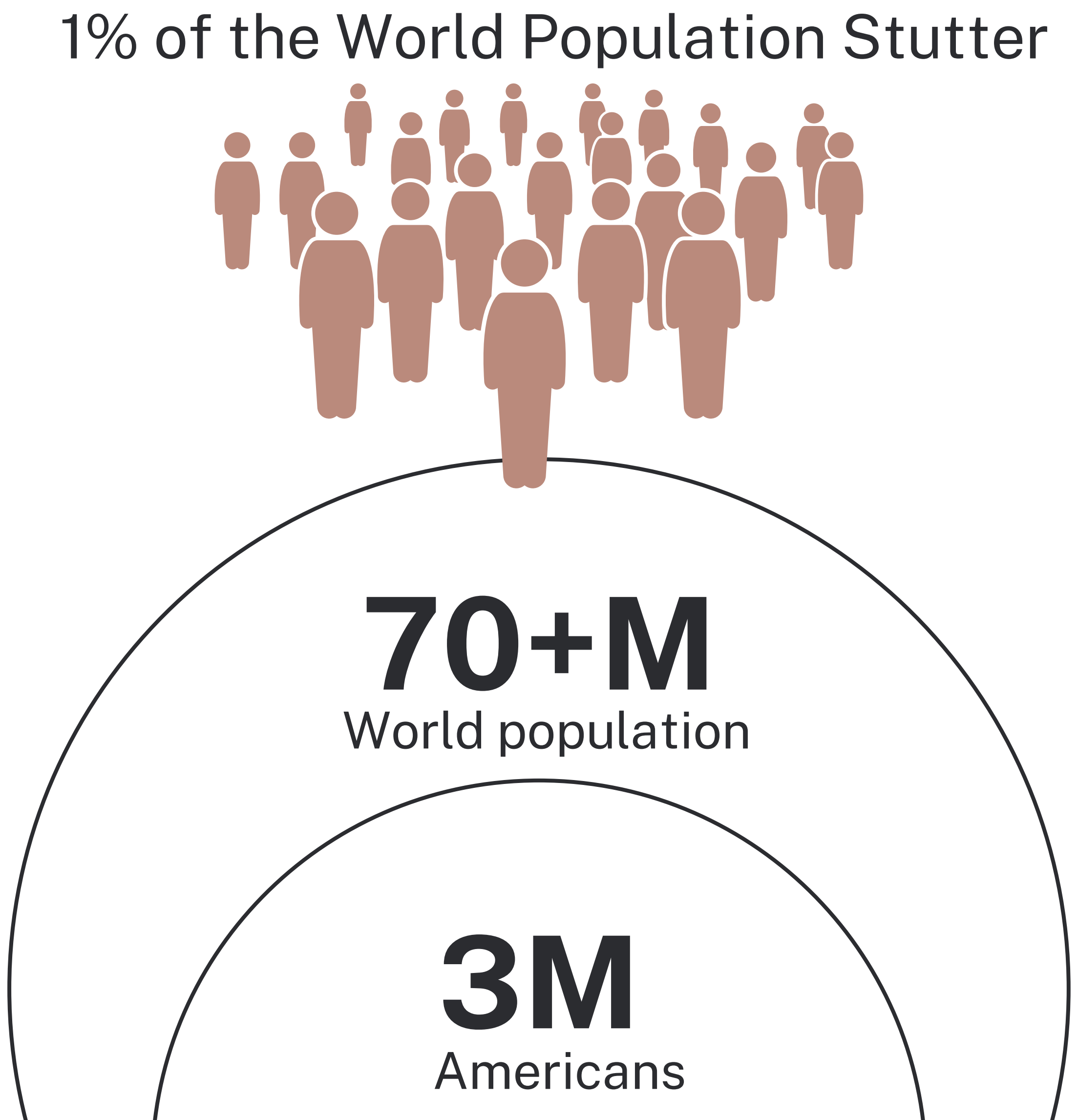
Speech AI relies heavily on human data — but that power is not equally distributed.

- Data are generally harvested and controlled by a few corporations (Eubanks 2018).
- Marginalized communities — like people who stutter — are poorly represented in AI data or have control on how their data is used by AI.



Sensitivity around Stuttered Speech Data

- Speech data can reveal personal identity and other health-related information.
- Misuse or misrepresentation may expose individuals to stigma or unwanted visibility.
- Contributing voice data can feel both empowering and risky for marginalized groups.



Macro Contextual Differerences across the Pacific — US vs China

- **Policy landscapes**
  - U.S. privacy laws are fragmented and sector-specific.
  - China's Personal Information Protection Law (PIPL) presents broader consumer control but excludes government oversight.

Macro Contextual Differerences across the Pacific — US vs China

- **Policy landscapes**
  - U.S. privacy laws are fragmented and sector-specific.
  - China’s Personal Information Protection Law (PIPL) presents broader consumer control but excludes government oversight.
- **Stigma** around stuttering is even stronger in China (Ma et al., 2023).

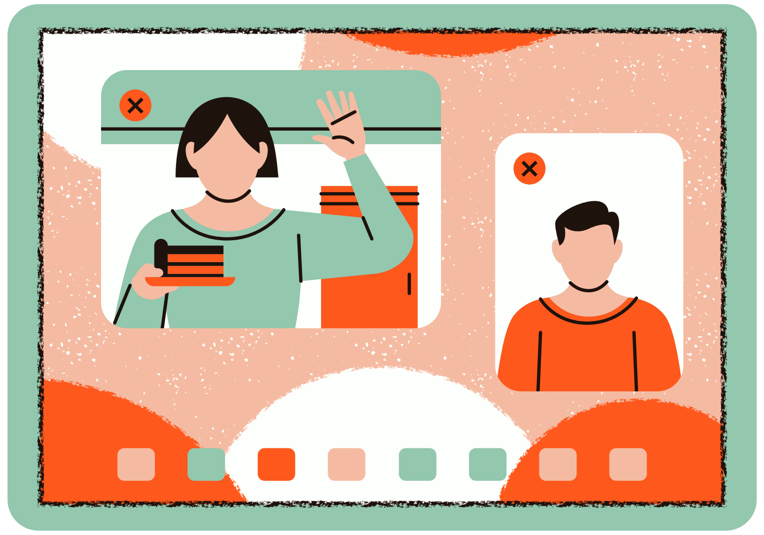
Macro Contextual Differerences across the Pacific — US vs China

- **Policy landscapes**
  - U.S. privacy laws are fragmented and sector-specific.
  - China’s Personal Information Protection Law (PIPL) presents broader consumer control but excludes government oversight.
- **Stigma** around stuttering is even stronger in China (Ma et al., 2023).
- **Data norms and expections**: e.g state surveillance more accepted by Chinese citizens (Kostka et al. 2021).

Research Questions

- RQ1: How can we use and govern disability data in an **ethical, respectful,** 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?

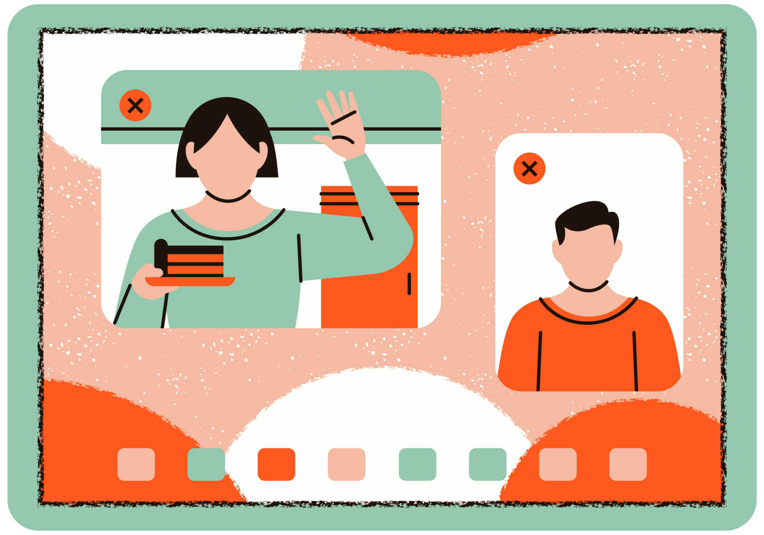
Methods



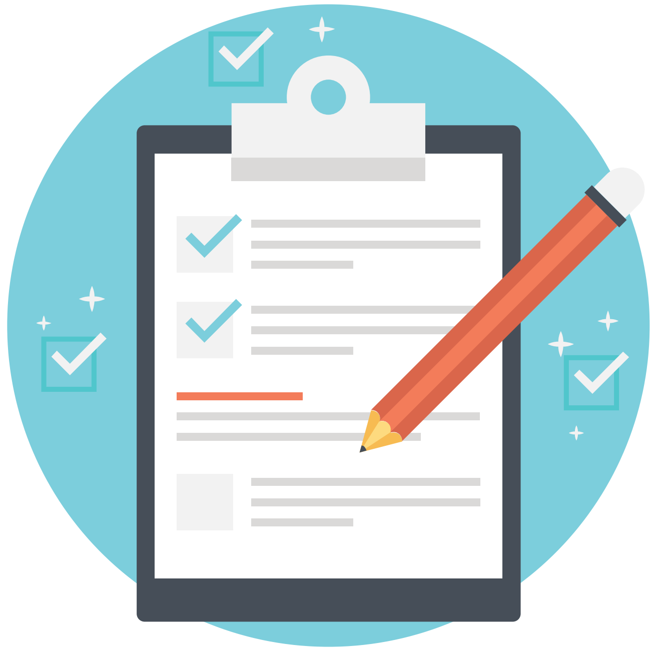
- **Interviews with 8 stuttering advocates**
    - 6 out of 8 self-identified as PWS
    - Deep dive in ethical challenges in speech data sharing, consent and ownership, long-term data stewardship.

Methods

AImpower.org



- **Interviews with 8 stuttering advocates**
  - 6 out of 8 self-identified as PWS.
  - Deep dive in ethical challenges in speech data sharing, consent and ownership, long-term data stewardship.



- **Survey with 149 community members**
  - 83 from China, 66 from the U.S.
  - Gather collective preferences on data-sharing motivations, governance preferences, trust.

Motivating Data Contribution: Transparency and Community Benefits

Communities want to contribute when goals are clear, community-centered, and transparent.

Motivating Data Contribution: Transparency and Community Benefits

Communities want to contribute when goals are clear, community-centered, and transparent.

> *‘ If it’s just framed as, ‘We’re collecting speech samples for X research project’ , it can feel unclear or unmotivating. If it’s presented as, ‘you have the chance to help create accessible AI for people who stutter — that’s exciting!’ ”*
>
> — Alex (stuttering community organizer and advocate, non-PWS)

Building Ethical Consent: Bridging Community Goals and Resource Constraints

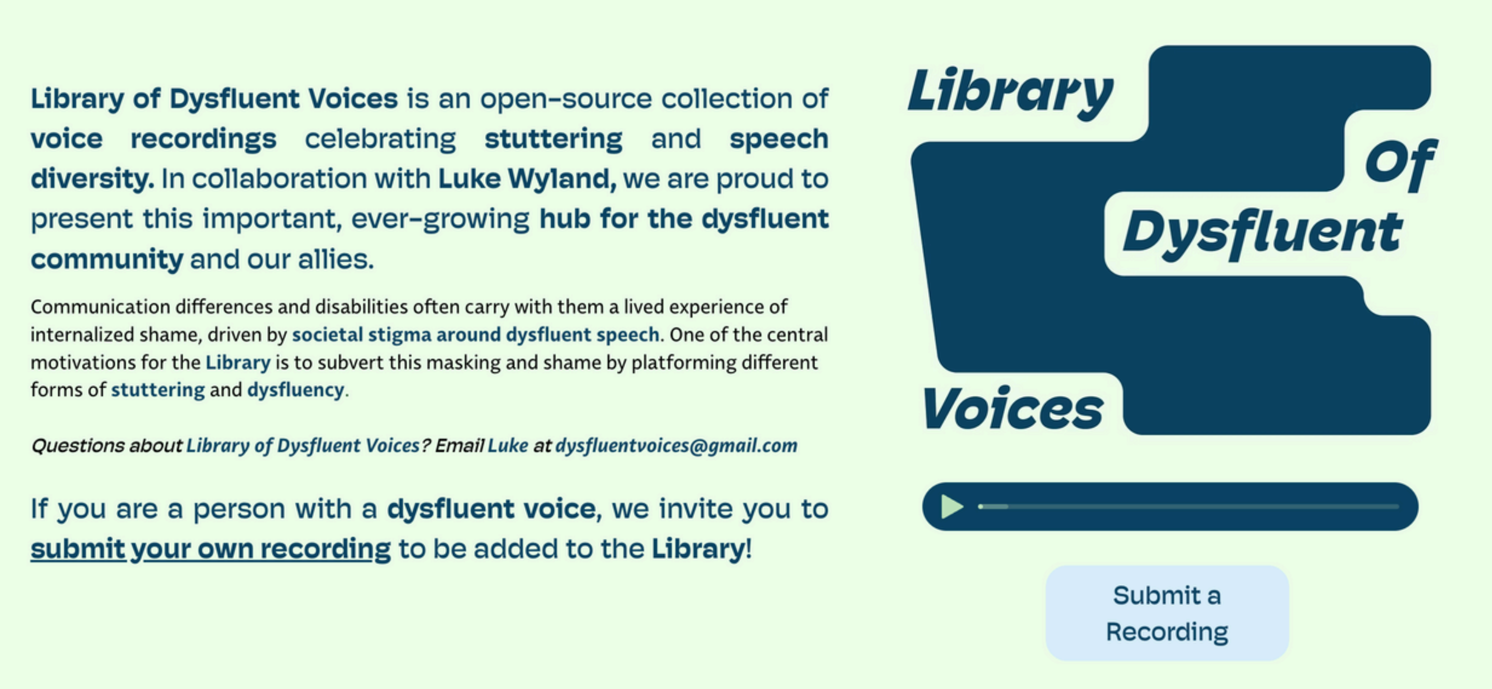
- Academic institutions utilize IRB: detailed procedures, benefits and risks, confidentiality, compensation, right to withdraw.

Building Ethical Consent: Bridging Community Goals and Resource Constraints

- Academic institutions utilize IRB: detailed procedures, benefits and risks, confidentiality, compensation, right to withdraw.
- Grassroots projects lack access and capacity for formal legal/governance infrastructure.
  - Rely on improvised consent forms: ethical misalignment & liability risk.

Building Ethical Consent: Bridging Community Goals and Resource Constraints

- Academic institutions utilize IRB: detailed procedures, benefits and risks, confidentiality, compensation, right to withdraw.
- Grassroots projects lack access and capacity for formal legal/governance infrastructure.
  - Rely on improvised consent forms: ethical misalignment & liability risk.



Library of Dysfluent Voices is an open-source collection of **voice recordings** celebrating **stuttering** and **speech diversity.** In collaboration with **Luke Wyland,** we are proud to present this important, ever-growing **hub for the dysfluent community** and our allies.

Communication differences and disabilities often carry with them a lived experience of internalized shame, driven by **societal stigma around dysfluent speech**. One of the central motivations for the **Library** is to subvert this masking and shame by platforming different forms of **stuttering** and **dysfluency**.

*Questions about **Library of Dysfluent Voices**? Email **Luke** at **dysfluentvoices@gmail.com***

If you are a person with a **dysfluent voice**, we invite you to **submit your own recording** to be added to the **Library**!

"We had to get some stock language up because we ran out of time. Most of it came from ChatGPT."

"I didn't think that at some point, a tech company might want to use these recordings for a different project."

— Alex

AImpower.org

Safeguarding Community Data: Open Access VS Privacy Risks

- Common safeguards: de-identification, easy mechanisms for requesting data deletion, transparent tracking of data use.

Safeguarding Community Data: Open Access VS Privacy Risks

- Common safeguards: de-identification, easy mechanisms for requesting data deletion, transparent tracking of data use.
- Password-protected or request-based access: Request and justify access, cite dataset creators, avoid unapproved uses.

Safeguarding Community Data: Open Access VS Privacy Risks

- Common safeguards: de-identification, easy mechanisms for requesting data deletion, transparent tracking of data use.
- Password-protected or request-based access: Request and justify access, cite dataset creators, avoid unapproved uses.

> *“We had a request to [use our data to] make an art installation... and it’s like no, that’s not why people gave us these recordings.”*
>
> — Natalie (SLP professor, data collector, non-PWS)

Safeguarding Community Data: Open Access VS Privacy Risks

- Common safeguards: de-identification, easy mechanisms for requesting data deletion, transparent tracking of data use.
- Password-protected or request-based access: Request and justify access, cite dataset creators, avoid unapproved uses.
- Significant barriers for cross-border data sharing due to legal complexity.

AImpower.org

Safeguarding Community Data: Protecting Identities in Voice Data Is Complex

- Advocates redact names & identifiers manually, but voices and stuttering patterns remain recognizable.
- De-identification is time-consuming and not robust for speech data.

Safeguarding Community Data: Protecting Identities in Voice Data Is Complex

- Advocates redact names & identifiers manually, but voices and stuttering patterns remain recognizable.
- De-identification is time-consuming and not robust for speech data.

> *“We manually removed names, but voices are still uniquely identifiable.”*
>
> — Teresa (stuttering community organizer, data collector, PWS)

Transparency Beyond Collection: Keeping Participants Informed

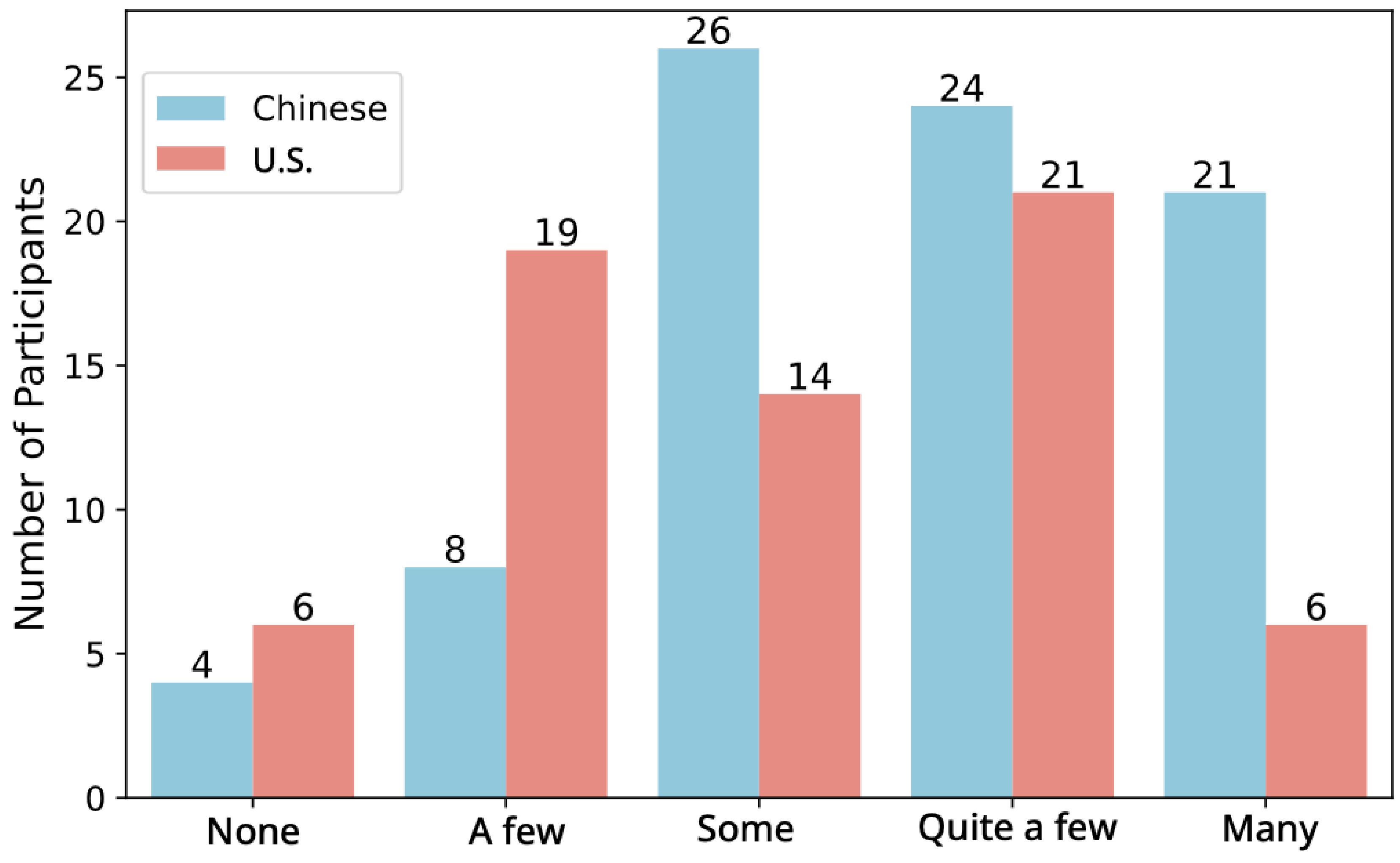
- Ongoing communication, not one-time consent.
- Updates about data usage, storage practices, and any secondary applications.
- Proactive communication channels: project websites or dashboards to share progress, milestones, and outcomes.

“I was never told how my data was used [in company-sponsored data collections].”

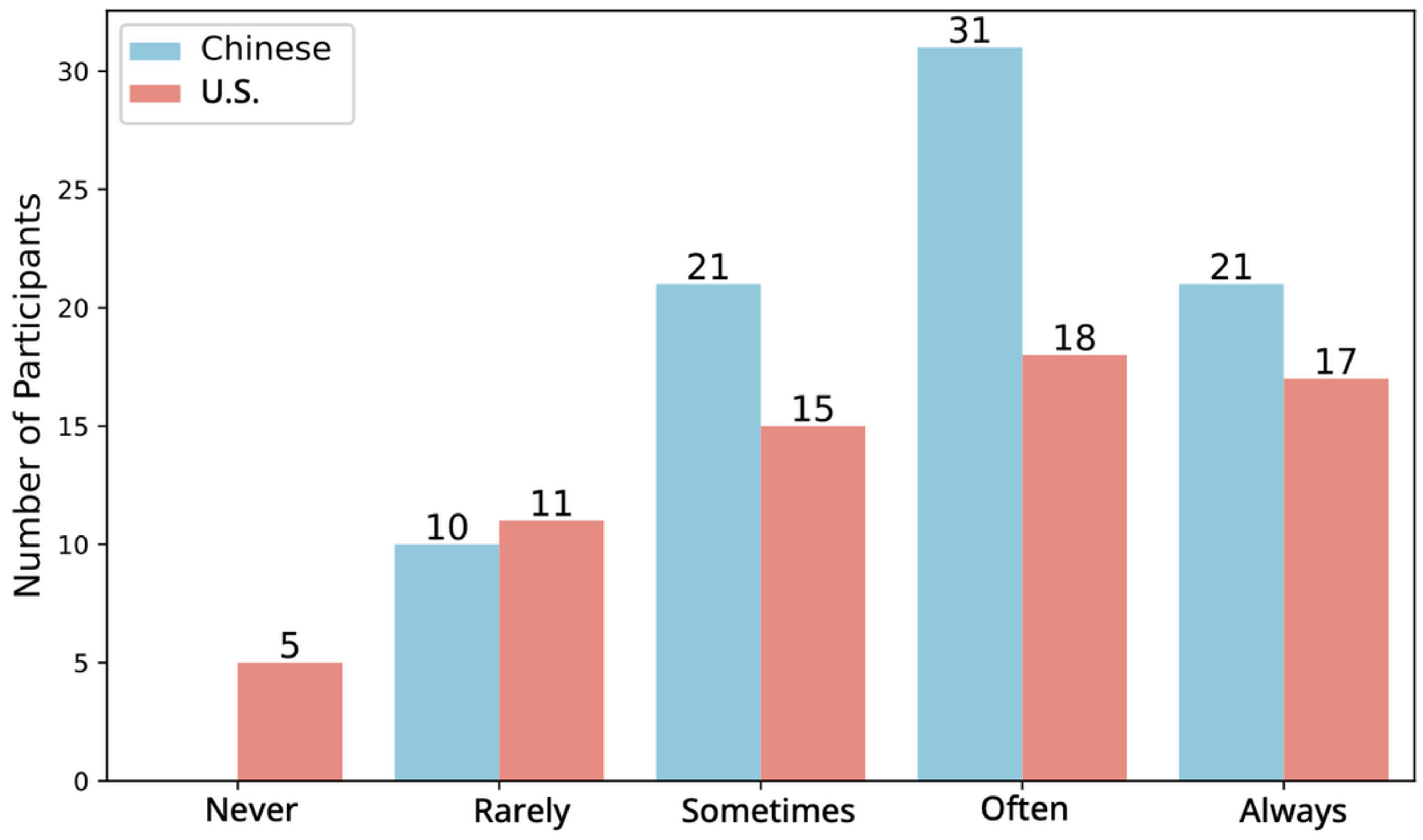
— Eric (stuttering community organizer and adovate, PWS)

Survey: Negative thoughts and attitude towards stutterng

More negative thoughts and greater avoidance among Chinese respondents.



"How often do you experience negative thoughts or feelings about your stuttering?"

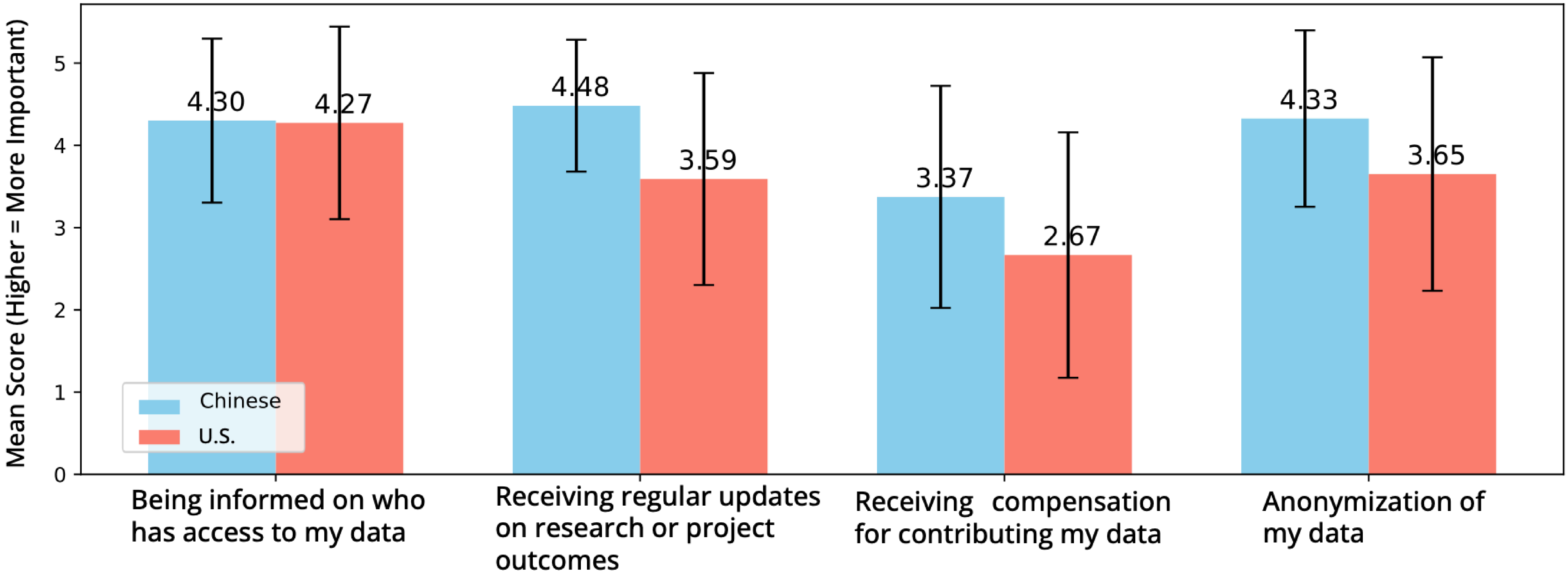


"How often do you try to avoid stuttering?"

AImpower.org

Survey: Factors influencing willingness to share data

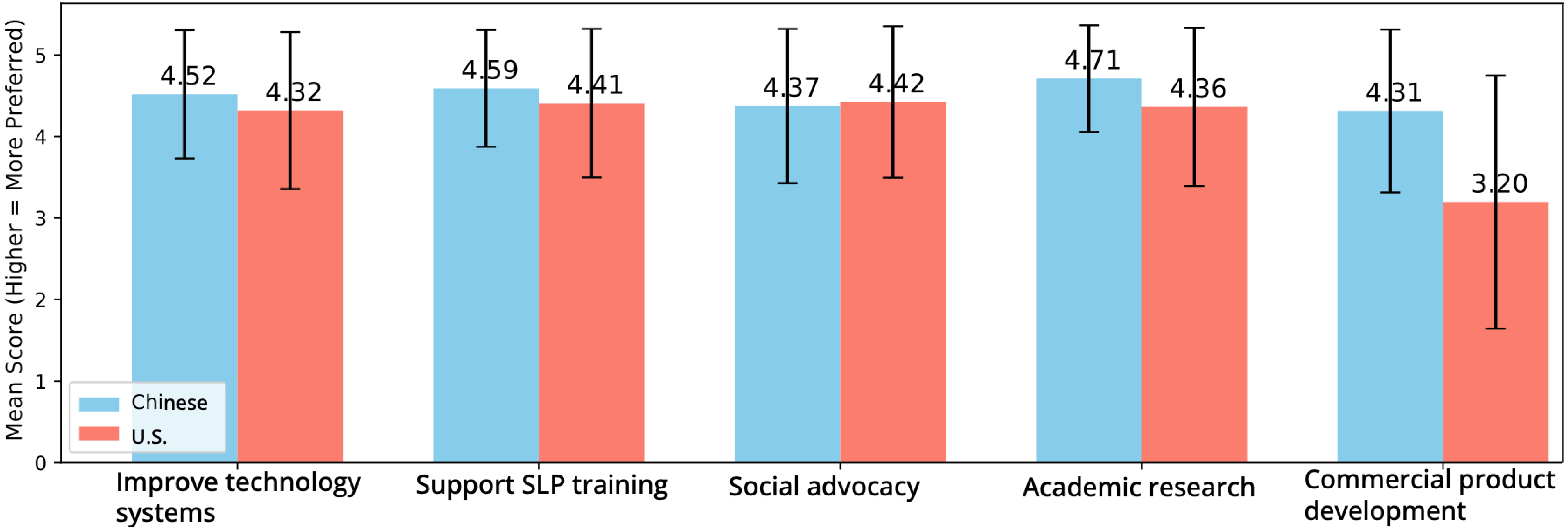
- Transparency around data access is equally valued by both groups.
- Receiving compensation was rated as the least important factor.



“Which of the following factors would encourage your willingness to share data?”

Survey: Acceptance ratings for different data-sharing purposes

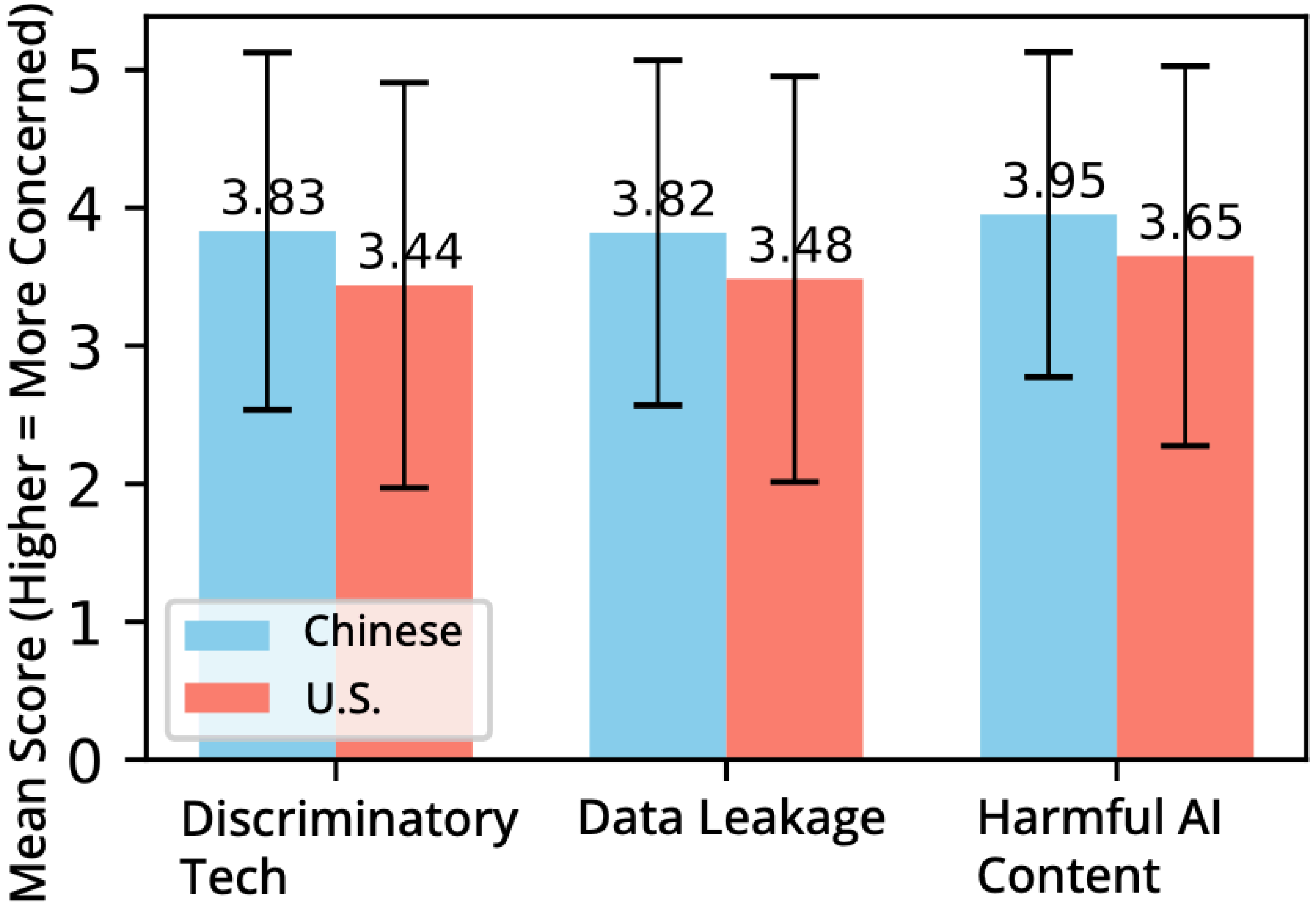
- Both are supportive of data sharing for public benefit.
- Chinese participants demonstrated significantly greater acceptance in academic research and commercial product development.



Preferences for different data use cases

Survey: Concerns About the Risks of Data Sharing

- Both shared moderate to high levels of concern across all three categories.

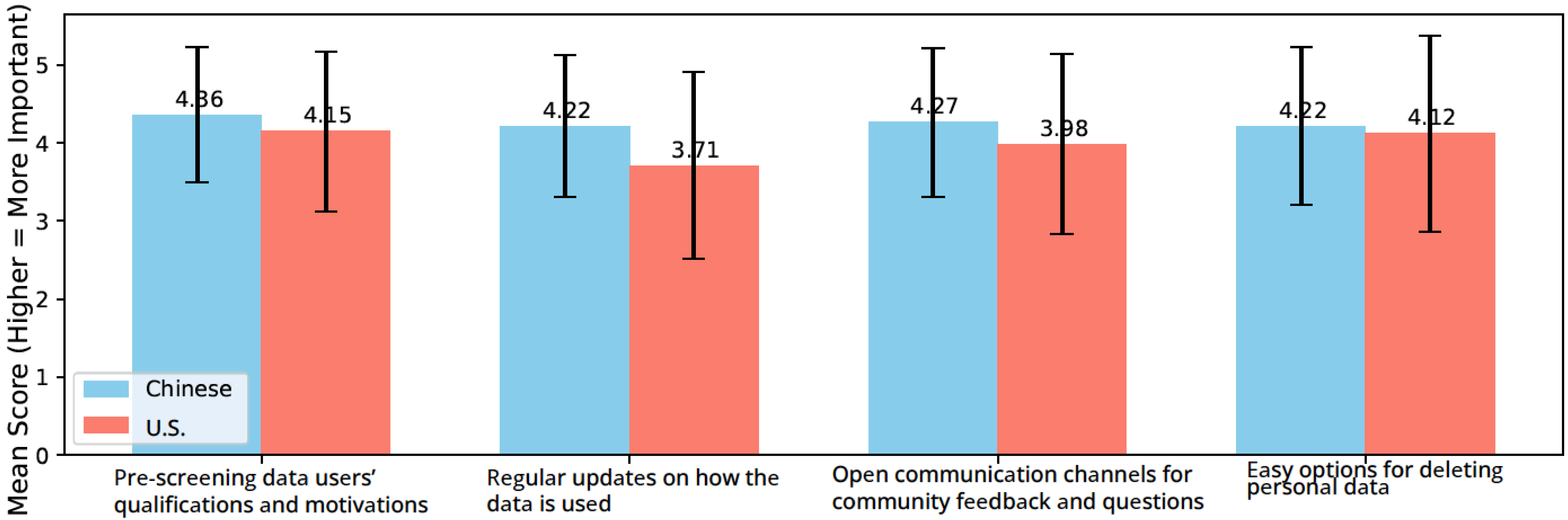


“How concerned are you about the following risks of data sharing?”

AImpower.org

Survey: Importance of four data protection and governance measures

- Both groups consider these measures as important.
- Chinese participants rated regular updates on data use significantly more important.



AImpower.org

Disucussion: Community-Centered Data Governance

- Governance is an ongoing relationship rather than a static contract.

Disucussion: Community-Centered Data Governance

- Governance is an ongoing relationship rather than a static contract.
- Moves governance from privacy protection to relational care, prioritizing trust and reciprocity beyond just access.

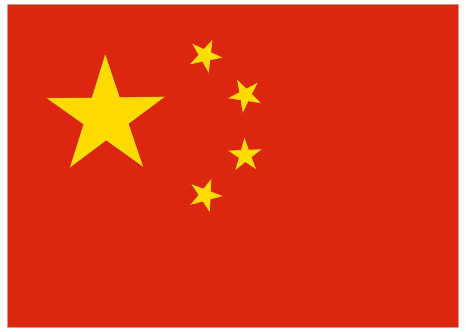
AImpower.org

Disucussion: Community-Centered Data Governance

- Governance is an ongoing relationship rather than a static contract.
- Moves governance from privacy protection to relational care, prioritizing trust and reciprocity beyond just access.
- Meaningful involvement of impacted communities across the full data lifecycle.

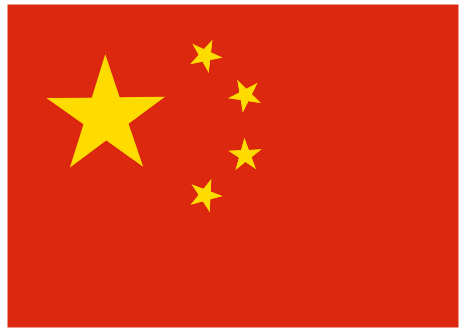
Disucussion: Universal Governance Values under Increasing AI Nationalism

- China: stricter data laws, higher social stigma
- U.S.: fragmented policy, active advocacy groups



Disucussion: Universal Governance Values under Increasing AI Nationalism

- China: stricter data laws, higher social stigma
- U.S.: fragmented policy, active advocacy groups
- Yet both communities:
    - Favor data use benefiting the stuttering community.
    - Shared concern on the use of data by commercial or media entities without community oversight or control.



AImpower.org

Future work

- Translating community values into **actionable, enforceable governance mechanisms**, such as consent structures, access protocols, and terms of use agreement.
- **Long-term contineous community engagement** to scaffold understanding of data use and governance: governance design involves complex trade-offs not fully captured in one or a few study sessions.

AImpower.org

Summary

- We explored **community preferences on how stuttered speech data should be collected, used, and governed** through interviews with stuttering advocates and a survey with PWS community members in both China and the United States.

- **Similarity in governance values** across both communities, emphasizing the universal importance of **transparency, trust, and agency in the stewardship of disability-related data,** despite substantial differences in how stuttering is perceived culturally.

Support our work

aimpower.org/support/

Contact us

Jingjin Li	**jingjin@aimpower.org**
Shaomei Wu	**shaomei@aimpower.org**

Read our paper

https://bit.ly/Stuttereddata