Maria Teleki

Research Statement

Socially Robust AI

I build methods, benchmarks, and audits for AI robustness in socially complex situations, at the intersection of Speech AI and Computational Social Science. My primary methodological move is a *diagnose-then-fix loop*: locate where AI commits a social failure – against a specific population (e.g., speakers who stutter) or a social construct from the social sciences (e.g., turn-taking rights) – measure it precisely enough to act on, then design targeted interventions to close the gap. My research connects three threads, each with short-term directions below and a shared long-term direction:

Disfluency-Aware AI: Is AI robust to how people naturally speak?

- Agents that read power and relationship to interrupt at the right moments
- Agents that produce disfluency deliberately, to shape what listeners retain

Personalized AI: Is AI robust to each person's uniqueness?

- Evaluation protocols for the $N{=}1$ user
- Personas grounded in lived experience via memories

Governable AI: Is AI robust within the sociopolitical structures it operates in?

- Audits for the full spectrum of governance interpretation

Long-Term Direction: I will build *socially situated reasoning*: social meaning is settled in context and no dataset can enumerate it, so the fix is reasoning at inference time. I will build methods that perform the social read, and accompanying benchmarks and audits.

My work spans speech venues (ICASSP, Interspeech), sociotechnical venues (FAccT, ICWSM), NLP venues (ACL, EMNLP Findings, LREC-COLING), and HCI venues (IUI, CHI) with co-authors from Cornell Tech, Google DeepMind, AAAS, KTH, Edinburgh, KIT, and ETH Zürich. I have also mentored twelve MS and undergraduate researchers, with nearly all co-authoring publications. My PhD is funded in part by the Avilés–Johnson Fellowship. The three threads below share one method, the *diagnose-then-fix* loop, applied to three layers of social context: how a person speaks, who they are, and the structures governing the interaction.

Disfluency-Aware AI: *Is AI robust to how people naturally speak?*

Voice is fast becoming a primary interface: voice mode in ChatGPT and Claude, smart glasses, and voice messages in iMessage and WhatsApp. And all of it runs on real, disfluent speech. As shown in Figure 1, dropping disfluency has a cost, and my first-authored work showed that ASR systems systematically fail to transcribe disfluencies (Interspeech '24 [13]),[1] so they never reach the corpora that SpeechLLMs learn from – making them out-of-distribution signals. The field's usual remedy, reinjecting disfluencies at random, treats as noise what carries dense, context-dependent social meaning: different genders use disfluencies differently (ICWSM '25 [14]), and disfluencies can be used to shape what listeners remember.[2] This results in worse downstream task performance, from spoken-content summarization (LREC-COLING '24 [12]) to conversational recommendation (Interspeech '25 [17]).

$\mathcal{Z}$-Scores (ICASSP '26 [15]) operationalizes Shriberg's disfluency taxonomy as a span-level metric, scoring how completely a system removes each category: interjections (*uh*, *um*), parentheticals (*I*

[1] Listeners filter disfluencies out of perception automatically, making them labor-intensive to annotate faithfully (Shriberg, 1994, *Preliminaries to a Theory of Speech Disfluencies*).

[2] Diachek & Brown-Schmidt, 2023, *The Effect of Disfluency on Memory for What Was Said*

mean), edited repairs (*Target omg sorry Walmart*). In a Switchboard case study, $\mathcal{Z}$-Scores pinpointed which disfluency categories a model handled worst; I designed a targeted prompt to address them, and those categories improved – a diagnose-then-fix loop that aggregate F1 can't support, because it never surfaces the per-category weakness in the first place. With collaborators at Cornell, I introduced the Linguistic Fidelity Spectrum (In Prep. for CSCW, [6]), pertaining to a deeper failure beneath disfluency removal itself: one utterance has many valid transcriptions, so scoring against a single reference measures only agreement with an arbitrary choice. I ship an open-source tool that hands transcript control to the user, and our follow-on work names this single-ground-truth assumption, reference monism (In Prep. for ARR [5]), as epistemic injustice in ASR evaluation.

With collaborators at KIT and ETH Zürich, I built Uh-Mazing (Under Submission at TACL [19]), a benchmark (to be released via Penn LDC) which measures disfluency-aware speech translation (Figure 1). With a collaborator at KTH, I built DRES (In Prep. for ARR [16]), an evaluation suite, which probes SpeechLLM backbones and finds stable editing policies shaped by training objective, with reasoning models systematically over-deleting fluent content. Finally, [Google DeepMind content redacted here.]

Figure 1: Smart glasses translate a patient's speech in real time. Because disfluencies signal a patient's confidence, translation that drops them loses clinically meaningful information, the gap our Uh-Mazing benchmark measures [19].

Short Term Direction. Real conversation is full-duplex — overlap, backchannels, interruptions that carry social meaning — but the field frames interruption as a timing or noise injection problem. Following [18], the missing layer is *social*: the same pause can invite interruption from a peer and discourage it from a subordinate.

Short Term Direction. My work so far asks whether systems are robust to human disfluency; the inverse is whether systems should produce it. Disfluencies shape listener memory and attention — a hesitation before a word makes it more memorable[2] — so a speaking agent can place them deliberately to emphasize information, flag uncertainty, or slow a listener down. I will build a system that generates communicative disfluency and run studies measuring what listeners retain, trust, and understand, turning disfluency into a signal that systems can use to better communicate with users.

Personalized AI: *Is AI robust to each person's uniqueness?*

Voice assistants, tutors, and agents are all racing to personalize. But adapting to a person first requires understanding them, and AI understands some people better than others: my work measures where that understanding breaks down, then closes the gap. I showed that discourse style alone shapes how well a person is represented: identical information is served unequally depending on speaking style (ICWSM '25 [14]). In SpeechLLMs the problem deepens, because voice and identity compound (Interspeech '26 [2]), and the harm is causal: voice conversion experiments show a speaker's voice directly shapes the trust, attribution, and empathy they receive (IUI Poster '26 [3]).

Short Term Direction. Building on [18], I will use mixed methods to build faithful evaluations for $N{=}1$ users the aggregate serves poorly – e.g., a rare set of medical conditions – where leave-one-out fails because there is no established construct of "good" to hold out against; the individuals do not generalize, but the protocol does.

Short Term Direction. When we build personas, we reduce a person to a few neat attributes (e.g.,

gender, race, age) but how someone reads a situation is generated in large part by a lifetime of accumulated context no attribute list can hold. Building on my storytelling survey (EMNLP Findings '25 [10]), memories in great literary works often come back in flashes of relevance, e.g., *At four, I fell off that swing in our backyard because I was staring at a flower, and it felt like slow motion. I remember my arm snapping when I hit the grass. My dad was so panicked, he switched to Hungarian, and couldn't switch back until we were well on our way to the ER.* That single fragment holds emotional, linguistic, and relational context a simple demographic slot discards. I will build memory-grounded models of lived experience, extending my persona work from the attributes a system assigns toward the history a person carries.

Governable AI: *Is AI robust within the sociopolitical structures it operates in?*

Governance is a moving target: what a policy means is situated, depending on who is affected and who enforces it. Auditing even one institution means AI reasoning jointly over policy documents, appeals procedures, and more; keeping that current across a sector is a standing computational commitment. Our FAccT '26 work [11] is the first instantiation. ACAI-US79 makes the divergence between two independent readers – a human panel recruited via Prolific and an LLM judge – its measurement object: the panel anchors the judge, so systematic human–model divergence reads as signal about the policy. Evaluations populate a scorecard over four dimensions (policy clarity, faculty support, feedback mechanisms, detection-tool governance) that aggregate into ACAI_u, an interpretable score in $[0, 100]$. I released the whole system – dataset, audit instrument, and a public site at `acai-us79.org` – so institutions can inspect their scores and others can extend the audit.

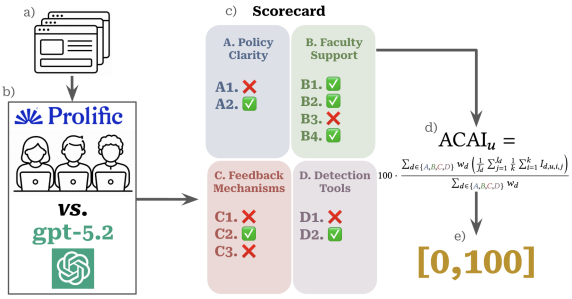

Figure 2: The ACAI-US79 pipeline: (a) AI policies are collected across 79 U.S. universities; (b) a human panel recruited via Prolific and an LLM judge each independently evaluate the policies; (c) evaluations populate a scorecard across four dimensions; (d) dimension scores aggregate into ACAI_u; (e) yielding an interpretable institutional score in $[0, 100]$.

Short Term Direction. No rule has a single reading; in ACAI-US79 the divergence between human and LLM readings was itself the finding. I will build audits that model the full spectrum of interpretation: a diverse population of personas reading each rule, using the persona infrastructure from my personalization work.

Long-Term Direction: Socially Situated Reasoning

My three threads converge on one bet: social meaning is situated, so AI serving people well is an inference-time reasoning problem.[3] What a word means, what a silence permits, what a policy requires is settled in context – *by who is speaking, to whom, with what lived experience, in what power dynamic, under which societal structures, at which point in time, and more.* Because this space of configurations is combinatorially too large for any dataset to enumerate, the social read has to happen at inference time via reasoning.

That reasoning can run over what my three threads have built: how a person speaks, who they are, and the structures governing the interaction – and the toolkit for it is maturing, mapped in my

[3] The situated-action tradition (Suchman, 1987, *Plans and Situated Actions*).

surveys of inference-time methods [7] and of narratives (central to meaning-making) [10]. And that reasoning has space to expand into other modalities. [Google DeepMind content redacted here.]

My agenda generalizes this pattern into a research program: benchmarks that score whether a system socially reads the situation correctly, methods that perform the reading, and audits that hold deployed systems accountable for it. While we should resist having our human lives reduced to a few neat social attributes, we also deserve to be understood and served well by AI [18]. Finding that middle ground, and moving the field toward it, is my long-term agenda.

Funding Landscape. I will aim for these funding opportunities: *Disfluency-Aware AI* and *Personalized AI* fit NSF CISE Robust Intelligence, Human-Centered Computing, and Information Integration and Informatics, as well as NSF SaTC 2.0, and Google and Amazon industry awards, and calls from the Cooperative AI Foundation (CAIF). *Disfluency-Aware AI* also fits the NIH; *Governable AI* fits the Knight Foundation and the Mozilla Foundation.

Students and Lab. Robust social reasoning is interdisciplinary, and so is the lab I want to build. My problems draw on CS/EE, information science, linguistics, psychology, and law. My pipeline moves students from contributing to my projects toward leading their own (detailed more in my Teaching Statement) [1, 9, 4, 8]; two projects earned first-place Student Research Week awards; one student completed an REU at RIT and three interned at Google, Meta, and Target this summer; a graduated undergraduate just started at Cornell Tech and graduated MS students have gone to Amazon and NVIDIA. I target venues across speech AI, computational social science, and NLP, making my lab a natural fit for many different types of students.

References

[1] Stephanie Birkelbach, Maria Teleki, Nehul Bhatnagar, Peter Carragher, Xiangjue Dong, and James Caverlee. SocialPulse: An Open-Source Subreddit Sensemaking Toolkit. In *SocialLLM @ ICWSM '26*, 2026.

[2] Shree Harsha Bokkahalli Satish, Christoph Minixhofer, Maria Teleki, James Caverlee, Ondrej Klejch, Peter Bell, Gustav Eje Henter, and Éva Székely. The Voice Behind the Words: Quantifying Intersectional Bias in SpeechLLMs. In *INTERSPEECH '26*, 2026.

[3] Shree Harsha Bokkahalli Satish, Maria Teleki, Christoph Minixhofer, Ondrej Klejch, Peter Bell, and Éva Székely. Walk a Mile in My Voice: Voice Conversion Shapes Trust, Attribution, and Empathy in Human–AI Speech Interactions. In *IUI '26 (Poster)*, 2026.

[4] Rohan Chaudhury, Maria Teleki, Xiangjue Dong, and James Caverlee. DACL: Disfluency Augmented Curriculum Learning for Fluent Text Generation. In *LREC-COLING '24*, 2024.

[5] Anna Seo Gyeong Choi, Maria Teleki, James Caverlee, Miguel del Rio, Corey Miller, and Hoon Choi. Beyond Single Ground Truth: Reference Monism as Epistemic Injustice in ASR Evaluation. In *In Prep. for ARR*, 2026.

[6] Anna Seo Gyeong Choi*, Maria Teleki*, Miguel del Rio, James Caverlee, Corey Miller, and Allison Koenecke. SpeechSpectrum: A Framework for Speech-to-Text Representation Along the Linguistic Fidelity Spectrum. In *In Prep. for CSCW*, 2027.

[7] Xiangjue Dong*, Maria Teleki*, and James Caverlee. A Survey on LLM Inference-Time Self-Improvement. In *arXiv*, 2024.

[8] Anna Irmetova, Haoran Liu, Maria Teleki, Peter Carragher, and James Caverlee. PodChecker: An Interpretable Fact-Checking Companion for Podcasts. In *MisD @ ICWSM '26*, 2026.

[9] Jason Kim, Maria Teleki, and James Caverlee. PromptHelper: A Prompt Recommender System for Encouraging Creativity in AI Chatbot Interactions. In *CHI '26 (Poster)*, 2026.

[10] Maria Teleki, Vedangi Bengali, Xiangjue Dong, Sai Tejas Janjur, Haoran Liu, Tian Liu, Cong Wang, Ting Liu, Yin Zhang, Frank Shipman, and James Caverlee. A Survey on LLMs for Story Generation. In *EMNLP Findings '25*, 2025. Extended Version Is Under Submission at TIST.

[11] Maria Teleki, Anna Seo Gyeong Choi, Anne Duray, Haoran Liu, Julie Zhang, Xiangjue Dong, Dilma Da Silva, Allison Koenecke, and James Caverlee. The Due Process Deficit: Auditing AI Governance in U.S. Higher Education. In *FAccT '26*, 2026.

[12] Maria Teleki, Xiangjue Dong, and James Caverlee. Quantifying the Impact of Disfluency on Spoken Content Summarization. In *LREC-COLING '24*, 2024.

[13] Maria Teleki, Xiangjue Dong, Soohwan Kim, and James Caverlee. Comparing ASR Systems in the Context of Speech Disfluencies. In *INTERSPEECH '24*, 2024.

[14] Maria Teleki, Xiangjue Dong, Haoran Liu, and James Caverlee. Masculine Defaults via Gendered Discourse in Podcasts and Large Language Models. In *ICWSM '25*, 2025.

[15] Maria Teleki, Sai Janjur, Haoran Liu, Oliver Grabner, Ketan Verma, Thomas Docog, Xiangjue Dong, Lingfeng Shi, Cong Wang, Stephanie Birkelbach, Jason Kim, Yin Zhang, and James Caverlee. Z-Scores: A Metric for Linguistically Assessing Disfluency Removal. In *ICASSP '26*, 2026.

[16] Maria Teleki, Sai Janjur, Haoran Liu, Oliver Grabner, Ketan Verma, Thomas Docog, Xiangjue Dong, Lingfeng Shi, Cong Wang, Stephanie Birkelbach, Jason Kim, Yin Zhang, Éva Székely, and James Caverlee. Conversational Speech Reveals Systematic Robustness Failures in LLMs. In *In Prep. for ARR*, 2026.

[17] Maria Teleki, Lingfeng Shi, Chengkai Liu, and James Caverlee. I want a horror – comedy – movie: Slips-of-the-Tongue Impact Conversational Recommender System Performance. In *INTERSPEECH '25*, 2025.

[18] Maria Teleki*, Kimi V. Wenzel*, Anna Seo Gyeong Choi*, Tobias Weinberg*, Shree Harsha Bokkahalli Satish*, Stephanny Sanchez*, Belu Ticona*, Ariadna Sanchez*, Yash Sonkar*, Aarti Mathur*, Christoph Minixhofer*, Abraham Glasser*, Raja Kushalnagar*, James Caverlee*, Minha Lee*, Shaomei Wu*, Alyssa Hillary Zisk*, Éva Székely*, Dylan Gaines*, Angelika Seeschaaf Veres*, Seray Ibrahim*, Nicholas Cummins*, and Allison Koenecke*. A Cross-Community Agenda for Speech AI. In *Under Submission at ARR*, 2026.

[19] Maike Züfle*, Maria Teleki*, Vilém Zouhar, Fabian Retkowski, Oliver Grabner, Alex Waibel, Jan Niehues, and James Caverlee. The Role of Disfluencies in Speech Translation. In *Under Submission at TACL*, 2026.