Research Statement

Tzu-Sheng Kuo

1 Overview

AI is transforming society's digital infrastructure, impacting vital decisions from how public agencies allocate services to how online communities moderate content. But who gets to shape this technology? **I envision a future where AI is shaped by the people it serves**, who best understand their own context, rather than relying on external developers who lack the domain knowledge to address local needs [1]. **I formalize this vision as *community-centered AI*** [2]. As a **human-computer interaction (HCI) researcher** combining system building, participatory design, and responsible AI, I build toward this vision by **creating interactive systems that empower people to harness their collective intelligence, aligning AI** with their needs, cultural values, and lived experiences.

My research is highly interdisciplinary and has impact across academia, industry, and policy. My first-author publications at top HCI and AI venues have earned a **CHI Best Paper Award** and a **NeurIPS Spotlight**. I have been named both a **Rising Star in ML & Systems** and a **Rising Star in Data Science**. At CMU, I was selected as the **Presidential Fellow in Ethics & Computational Technologies** and **inaugural Fellow on AI & Society**. I have **secured $117,000 USD as a PI or Co-PI** in research grants during my PhD, bridged disciplines by organizing workshops at ICML, ACL, and CHI, and collaborated with scholars across three continents. Beyond academia, I work closely with industry partners like Microsoft and Google DeepMind, civic initiatives like MetaGov and the Wikimedia Foundation, and local communities from Pittsburgh to Taiwan. Finally, my research informs a policy report from the **European Parliament** and consultations with the **White House Office of Science & Technology Policy**.

2 Empowering People to Shape AI Collaboratively

While people seek more agency to shape the AI deployed in their communities, realizing the vision of *community-centered AI* is challenging in practice [2]. In particular, members of the same community frequently have conflicting views on how their AI should be designed. Even when they reach consensus, they often lack the technical expertise to implement their decisions. My PhD overcomes these challenges through systems that surface disagreements, facilitate deliberation, and translate collective decisions into AI artifacts. Below, I showcase three such systems, enabling groups of *non-AI experts* to *collaborate* on shaping AI policies, agents, and datasets [3, 4, 5].

2.1 PolicyCraft: Collaborative Policy Design for AI Governance

Problem: Today, top-down policy development often yields policies misaligned with local needs and lacking legitimacy. For example, the Dutch government one-sidedly deployed a policy using AI for childcare benefits, which ignored thousands of families' daily struggles and led to the government's resignation. Seeking to avoid such disasters, civic tech enthusiasts have created tools like Polis and Remesh that solicit bottom-up input. But these tools still limit citizens to giving feedback, leaving policymakers with the final say on how policy should be written. Empowering people to actively co-author policy has the potential to address policy misalignment, but doing so is challenging: because policy language is high-level and abstract, citizens can struggle to translate their on-the-ground experiences into policy drafts. This abstraction also masks underlying disagreements: people may agree on a policy's wording, only to find their interpretations diverge when applying it to real-world cases that the policy is meant to govern.

Research: To address these challenges, I developed PolicyCraft [3], a system that empowers people to co-author policies through a novel workflow I call *case-grounded deliberation*. This workflow aligns high-level policies with people's stances on real-world cases. First, people share cases from their lived experiences and discuss how they should be handled, regardless of what current policies say. Upon reaching consensus on a case, they use it to critique and revise current policies, or propose a new one. Here, AI augments the case-to-policy translation: for example, when a case reveals a policy flaw, the system suggests potential revisions. Together, this workflow bridges the gap between abstract policies and concrete cases, and helps people pinpoint the source of their disagreements: whether they disagree on a policy's wording, or fundamentally disagree on how a case should be handled. In a field study of university students co-authoring AI course policies, students using PolicyCraft reached greater consensus and developed better-supported policies, compared with those using a baseline system without case-grounded support.

Impact: Two CMU courses adopted policies students co-authored with PolicyCraft, inspiring student-led, campus-wide AI guidelines at the University of Maryland, Baltimore County. Beyond classrooms, I used PolicyCraft with MetaGov to design organizational policies, and applied the case-grounded approach to help gig workers and policymakers co-design platform policies [6, 7]. The European Parliament's Policy Department for Justice, Civil Liberties, and Institutional Affairs highlighted PolicyCraft as an exemplar of using AI for public deliberation on EU law.

2.2 Botender: Collaborative Design of AI Agent Behavior

Problem: Online communities rely on bots for diverse tasks like content moderation. Yet, off-the-shelf bots built by outsiders often carry assumptions about how communities should operate that clash with their needs. For example, in a peer-support community for mental health, a pre-built bot might remove posts about self-harm, silencing vulnerable members seeking help. As an alternative, LLMs allow communities to design custom AI agents by writing prompts. Because prompting requires no coding, members can do so together, pooling their insider knowledge to anticipate myriad scenarios their agent must handle. However, they face two challenges: First, as non-AI experts, they often tweak a prompt with a few scenarios in mind, unaware their edits may trigger undesirable agent behavior in other scenarios. Second, they may agree on a prompt's wording (e.g., "keep the community safe"), only to discover after deployment that they interpret it differently (e.g., what "safe" means), and thus expect different agent behaviors. These challenges risk communities deploying harmful agents that alienate members and damage community bonds.

Research: To address these challenges, I developed Botender [4], a system that enables online communities to collaboratively design, iterate on, and deploy custom LLM agents directly within their platforms. In Botender, I introduced *case-based provocations*: AI-simulated scenarios of human-agent interactions designed to spark reflection and discussion. These scenarios are dynamically generated to reveal pitfalls in prompts written by non-AI experts, such as ambiguous or overly narrow phrasing, that cause unintended agent behavior in scenarios they had not considered. In turn, these scenarios anchor people's discussions in concrete examples of the agent's behavior beyond the prompt's wording, helping them align on how the agent should act before it goes live. Our lab study showed that, unlike standard test cases that mostly validate expected agent behavior, case-based provocations better surface disagreements and expose prompt flaws. Our field deployment across six Discord servers demonstrated that Botender effectively enabled real-world communities to design custom agents tailored to their unique contexts.

Impact: The Discord communities that participated in our field deployment have chosen to continue using Botender after the study's conclusion. Beyond online communities, case-based provocations have been applied to high-stakes domains like mental health and law [8]. Furthermore, I am extending this approach to support large-scale public participation in collectively designing AI constitutions, as detailed in my future agenda.

2.3 Wikibench: Collaborative Data Curation for AI Evaluation

Problem: AI tools are increasingly deployed in community contexts, but the benchmarks used to evaluate them are typically curated by AI developers, who pay crowdworkers outside the community to label data. Because outsiders label data without knowing local contexts, this knowledge gap threatens AI evaluation. For example, research shows that crowdsourced datasets systematically label harmless African American English (AAE) phrases as toxic. If such datasets were used to evaluate a content moderation AI in an AAE-speaking community, the evaluation could not tell when the AI falsely censors everyday language, as the "ground truth" provided by crowdworkers sees it as correct behavior. While having communities curate their own datasets would close this knowledge gap, doing so is challenging in practice. Even within a community, members can disagree over how a data point should be labeled, or how labels should be defined. Resolving these disagreements is hard as datasets scale and people work asynchronously.

Research: To address these challenges, I developed Wikibench [5], a system that enables communities to collaboratively curate AI datasets. For this project, I partnered with the Wikipedia community and integrated Wikibench directly into their platform, unlike typical data labeling tools hosted on standalone sites. This deep integration embeds data curation into members' everyday workflows without overhead, resolving the asynchronous friction of external sites that offer no incentive for sustained engagement. Furthermore, Wikibench computationally surfaces disputed items and notifies members about them—whether individual data points, label definitions, or the overall data curation plan—directing attention where needed most. Ultimately, Wikibench embodies the vision of *community-centered AI* [2], where the curation process is led and governed by Wikipedians themselves. Our analysis shows that datasets curated via Wikibench align better with Wikipedia's content moderation norms than data labeled by individuals on an external site. As one Wikipedian said of Wikibench, *"It was meant to be there."*

Impact: The Wikimedia Foundation adopted Wikibench's dataset as a benchmark to evaluate their AI models, and we co-organized an ACL workshop with a featured track on Wikipedia datasets. Beyond Wikipedia, Wikibench's vision of community-centered AI directly informed the launch of Community Benchmarks on Kaggle, the world's largest data science platform. Finally, this project introduced a community-centered dimension to *data-centric AI*, a growing field that positions deliberative data curation as a first-class approach to improving AI [9, 10, 11].

3 Future Research Agenda

I envision AI systems shaped by and built for the people they serve. Drawing on my research across HCI and AI, I outline an agenda driven by two reciprocal directions, creating a virtuous cycle where collective *human* intelligence and *artificial* intelligence continually strengthen each other. I conclude with a concrete strategy to secure funding.

3.1 Harnessing Collective Intelligence to Shape AI

My research thus far has enabled people to collaborate on AI development, through systems that surface disagreements, facilitate deliberation, and translate collective decisions into AI artifacts. What remaining sociotechnical challenges must we overcome to fully empower people to shape AI, and what algorithmic opportunities might emerge?

Sociotechnical Challenges: One foundational challenge is developing the infrastructure required for participation at scale. For example, as democratic countries begin to develop sovereign AI built for their citizens, aligning it with the public interest requires scaling participation to the societal level. Future research must address the risks and bottlenecks of collaborative AI development amplified by this scale, such as mitigating information overload during collective sensemaking, preventing fragmented effort in collective action, and countering adversarial manipulation by malicious actors. Building upon PolicyCraft and Botender, I have begun tackling these challenges alongside the Public AI Network, MetaGov, and Change.org to build the participatory infrastructure for shaping the constitution of Apertus AI, an open-source model from the Swiss AI Initiative that emphasizes democratic participation. Our goal is to develop infrastructure and generate insights that will generalize across broader civic and institutional contexts.

Algorithmic Opportunities: Collaborative AI development also opens new directions for model training and alignment. For example, I propose complementing the current ML paradigm of *learning from demonstration* with what I term *learning from deliberation* [2]. Today's ML often relies on explicit feedback, labels, or human actions to guide training. But in collaborative contexts where deliberative traces like exchanged opinions already exist, requiring users to provide artificial labels introduces overhead. For instance, rather than requiring Wikipedians to label data using Wikibench after their discussion has already concluded, AI could directly leverage their organic discussions for alignment. This not only reduces human workload, but also enables AI to learn the latent collective intelligence unavailable in explicit labels [10]. Yet, realizing this requires algorithms that weigh relative importance and distill valuable signals from noisy discourse, while preserving community agency over what AI learns.

3.2 Developing AI to Augment Collective Intelligence

Conversely, how can we develop AI systems that actively augment people's collective intelligence? For example, Botender's AI-generated *case-based provocations* enable users to collectively reflect on the desiderata of community governance. I plan to expand this line of work across new contexts, from the future of work to civic participation.

AI for Future of Work: Organizations and workplaces are rapidly adopting AI across teams, yet today's AI tools, from coding assistants to writing tools, are primarily designed for solo work. I plan to develop AI systems that reduce friction and enhance synergy in collaborations to foster team success and productivity. For example, team members often think they have reached a conclusion at the end of a meeting, only to realize later that their understandings diverge. How can we design AI agents to proactively surface these hidden misalignments and keep collective efforts aligned? Beyond reducing friction, how can AI prevent team failure through simulation, facilitate new collaborations, or foster brainstorming to accelerate innovation? I will pursue these directions by drawing on my existing work, including Botender's case-based provocations as a *tool for thought*, as well as my broader research on AI agent simulation [12], AI group facilitation [13], and AI's impact on creativity [14].

AI for Civic Participation: Finally, I am passionate about leveraging AI to broaden and deepen civic participation. Today's democratic mechanisms are largely inaccessible to everyday citizens who lack the time and knowledge to attend a 2PM town hall or read 100-page policy reports. This leaves public decisions disproportionately influenced by well-resourced interest groups. Can we develop AI-augmented civic technologies that equip citizens with the knowledge required to participate, remove barriers to civic input and deliberation, and empower the public to actively track and shape policy outcomes? I will build upon my research on PolicyCraft and draw from my experience empowering unhoused individuals to shape homeless services [1, 15]. I aim to pursue these research directions not through the dominant framing of using AI for *efficiency*, but instead with the goal of strengthening the *resilience* of democratic mechanisms with a focus on equitable representation, deliberative depth, and procedural legitimacy.

3.3 Target Funding Sources

Building on my record of securing $117k USD, I will sustain my research with federal, industry, and civic funding. My HCI and community-centered research aligns with NSF CISE's Human-Centered Computing (HCC) and Connected Communities (CPS) programs. I will also leverage my industry connections to pursue grants like Microsoft's New Future of Work and Google's AI for Organizations. In the civic space, my research aligns with the growing public discourse and funding around AI regulation. I will seek foundation grants from Mozilla, PIT-UN, and Humanity AI.

References

[1] **Tzu-Sheng Kuo**, Hong Shen, Jisoo Geum, Nev Jones, Jason I. Hong, Haiyi Zhu, and Kenneth Holstein. "Understanding Frontline Workers' and Unhoused Individuals' Perspectives on AI Used in Homeless Services". *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. CHI '23. **Best Paper Award (Top 1%)**.

[2] **Tzu-Sheng Kuo**, Quan Ze Chen, Amy X. Zhang, Kenneth Holstein, and Haiyi Zhu. "Community-Centered AI is Feasible and Beneficial for Impacted Communities". *Advances in Neural Information Processing Systems — Position Paper Track*. NeurIPS '26. **Spotlight (Top 2%)**.

[3] **Tzu-Sheng Kuo**, Quan Ze Chen, Amy X. Zhang, Jane Hsieh, Haiyi Zhu, and Kenneth Holstein. "PolicyCraft: Supporting Collaborative and Participatory Policy Design through Case-Grounded Deliberation". *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. CHI '25.

[4] **Tzu-Sheng Kuo**, Sophia Liu, Quan Ze Chen, Joseph Seering, Amy X. Zhang, Haiyi Zhu, and Kenneth Holstein. "Botender: Supporting Communities in Collaboratively Designing AI Agents through Case-Based Provocations". *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems*. CHI '26.

[5] **Tzu-Sheng Kuo**, Aaron Lee Halfaker, Zirui Cheng, Jiwoo Kim, Meng-Hsin Wu, Tongshuang Wu, Kenneth Holstein, and Haiyi Zhu. "Wikibench: Community-Driven Data Curation for AI Evaluation on Wikipedia". *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. CHI '24.

[6] Jane Hsieh, Angie Zhang, Sajel Surati, Sijia Xie, Yeshua Ayala, Nithila Sathiya, **Tzu-Sheng Kuo**, Min Kyung Lee, and Haiyi Zhu. "Gig2Gether: Datasharing to Empower, Unify and Demystify Gig Work". *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. CHI '25. **Best Paper Honorable Mention (Top 5%)**.

[7] Sander de Jong, Jane Hsieh, **Tzu-Sheng Kuo**, Rune Møberg Jacobsen, Niels van Berkel, and Haiyi Zhu. "Understanding, Challenging, and Demystifying Perceptions of Gig Worker Vulnerabilities". *Proceedings of the 5th Annual Symposium on Human-Computer Interaction for Work*. CHIWORK '26.

[8] K. J. Kevin Feng, **Tzu-Sheng Kuo**, Quan Ze Chen, Inyoung Cheong, Kenneth Holstein, and Amy X. Zhang. "PolicyPad: Collaborative Prototyping of LLM Policies". *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems*. CHI '26.

[9] Mark Mazumder, Colby Banbury, Xiaozhe Yao, Bojan Karlaš, William Gaviria Rojas, Sudnya Diamos, Greg Diamos, Lynn He, Alicia Parrish, Hannah Rose Kirk, Jessica Quaye, Charvi Rastogi, Douwe Kiela, David Jurado, David Kanter, Rafael Mosquera, Juan Ciro, Lora Aroyo, Bilge Acun, Lingjiao Chen, Mehul Smriti Raje, Max Bartolo, Sabri Eyuboglu, Amirata Ghorbani, Emmett Goodman, Oana Inel, Tariq Kane, Christine R. Kirkpatrick, **Tzu-Sheng Kuo**, Jonas Mueller, Tristan Thrush, Joaquin Vanschoren, Margaret Warren, Adina Williams, Serena Yeung, Newsha Ardalani, Praveen Paritosh, Ce Zhang, James Zou, Carole-Jean Wu, Cody Coleman, Andrew Ng, Peter Mattson, and Vijay Janapa Reddi. "DataPerf: Benchmarks for Data-centric AI Development". *Advances in Neural Information Processing Systems — Datasets and Benchmarks Track*. NeurIPS '23.

[10] Shachar Don-Yehiya, Ben Burtenshaw, Ramon Fernandez Astudillo, Cailean Osborne, Mimansa Jaiswal, **Tzu-Sheng Kuo**, Wenting Zhao, Idan Shenfeld, Andi Peng, Mikhail Yurochkin, Atoosa Kasirzadeh, Yangsibo Huang, Tatsunori Hashimoto, Yacine Jernite, Daniel Vila-Suero, Omri Abend, Jennifer Ding, Sara Hooker, Hannah Rose Kirk, and Leshem Choshen. "The Future of Open Human Feedback". *Nature Machine Intelligence* (2025).

[11] Luis Oala, Manil Maskey, Lilith Bat-Leah, Alicia Parrish, Nezihe Merve Gürel, **Tzu-Sheng Kuo**, Yang Liu, Rotem Dror, Danilo Brajovic, Xiaozhe Yao, Max Bartolo, William A Gaviria Rojas, Ryan Hileman, Rainier Aliment, Michael W. Mahoney, Meg Risdal, Matthew Lease, Wojciech Samek, Debojyoti Dutta, Curtis G Northcutt, Cody Coleman, Braden Hancock, Bernard Koch, Girmaw Abebe Tadesse, Bojan Karlaš, Ahmed Alaa, Adji Bousso Dieng, Natasha Noy, Vijay Janapa Reddi, James Zou, Praveen Paritosh, Mihaela van der Schaar, Kurt Bollacker, Lora Aroyo, Ce Zhang, Joaquin Vanschoren, Isabelle Guyon, and Peter Mattson. "DMLR: Data-centric Machine Learning Research - Past, Present and Future". *Journal of Data-centric Machine Learning Research* (2024).

[12] **Tzu-Sheng Kuo**, Carrie J. Cai, Meredith Ringel Morris, and Michael Terry. "Synthetic TLX: Forecasting Human Workload Using Agent Simulation". *Under Review for the 2027 CHI Conference on Human Factors in Computing Systems*. CHI '27.

[13] Ziyi Liu, **Tzu-Sheng Kuo**, Jiale Liu, Deqing Fu, Nithum Thain, and Crystal Qian. "Learning to GUIDE the Group: Self-Learning LLM Facilitation Across Multi-Party Tasks". *Under Review for the 15th International Conference on Learning Representations*. ICLR '27.

[14] Yulin Yu, Yizhou Li, **Tzu-Sheng Kuo**, Kentaro Toyama, Siddharth Suri, and Scott Counts. "From Searchable to Non-Searchable: Generative AI and Information Diversity in Online Information Seeking". *Under Review for the 2027 CHI Conference on Human Factors in Computing Systems*. CHI '27.

[15] Ningjing Tang, Jiayin Zhi, **Tzu-Sheng Kuo**, Calla Kainaroi, Jeremy J. Northup, Kenneth Holstein, Haiyi Zhu, Hoda Heidari, and Hong Shen. "AI Failure Cards: Understanding and Supporting Grassroots Efforts to Mitigate AI Failures in Homeless Services". *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*. FAccT '24.