

SANJAY KAIRAM

San Francisco, CA · skairam@cs.stanford.edu · [linkedin.com/in/skairam](https://www.linkedin.com/in/skairam) · scholar.google.com

SUMMARY

Data scientist and research leader with 15+ years of experience building measurement systems for human behavior, model quality, safety & governance within frontier AI and large-scale online sociotechnical systems. At OpenAI, served as the sole data scientist within the Research Org, building statistical methods and evaluation infrastructure for frontier model development, pre-release assessment, and launch decisions, including mixed-effects modeling for human-data campaigns, evaluator expertise modeling, and analysis of real-world AI coding agent behavior. Previously led research on safety measurement and moderator workflows at Reddit and Twitch, with public-facing research on proactive interventions, annotation quality, moderator workflows, and community outcomes (CHI, CSCW, ICWSM; [1,600+ citations](#)).

Focus areas: human evaluation & measurement systems · user and model behavior analysis · safety & harm measurement · governance and moderation · sociotechnical systems & computational social science

EXPERIENCE

OpenAI · San Francisco, CA · 2024–2026

Member of Technical Staff, Human Data — Research · Product

Sole data scientist in OpenAI's Research org, defining methodology for frontier model evaluation across post-training, alignment, and preparedness teams. Later served as Product DS lead for ChatGPT for Work, leading strategy and experimentation for Connectors and Company Knowledge.

- Pioneered **mixed-effects hierarchical modeling** for human-data evaluation campaigns at OpenAI, separating trainer, task, and campaign-design variance from true model signal; adopted as a methodological standard across multiple model-development teams.
- Designed a human evaluation protocol for **flagging novel behavioral regressions** in new model candidates: introduced an "authentic tone" dimension in blind paired comparisons that detected sycophantic behavior on an online A/B test had previously rewarded (e.g. sycophancy), demonstrating how targeted expert evaluations can complement user feedback in pre-launch model assessment.
- Supported **Codex research preview** with large-scale analysis of ~1M agent-authored pull requests scraped from OSS repos, using LLM-assisted classification of unstructured code, candidate-selection modeling in multi-candidate workflows, and rollout latency decomposition to inform post-training and evaluation priorities.
- Built **checkpoint-monitoring analysis infrastructure** for a mainline reasoning model pre-release training run, tracking progress in correctness and deception across math, science olympiad, and multidisciplinary benchmarks.
- Improved **human data campaign efficiency** through two complementary systems: designed and deployed a Bayesian mixed-effects trainer–task matching system live within OpenAI's human data platform, with offline evaluation showing **up to 6% improvement in first-attempt success rates**; and a 4-question Prolific evaluator screener achieving **97% recall / 46% precision** against downstream trainer campaign performance, providing substantial reduction in campaign screening costs.
- Led Product DS for ChatGPT for Work: led analytics and experimentation that **doubled adoption of Connectors over Q4** by identifying end-user linkage as a primary bottleneck, and then designing/analyzing **12 sequential A/B tests** on Connector Discoverability interventions, each driving **10-25% relative gains in linkage rates**, with several directly driving significant gains in Connector DAU.

Reddit · San Francisco, CA · 2022–2024

Head of Research Science, Community · Staff Research Scientist, Community, Moderation & Governance

Empirical lead for community governance, moderation-tooling evaluation, and platform-wide quality measurement; primary research partner for AI/ML-powered moderation and governance products; led Reddit for Researchers, connecting anonymized large-scale platform data to hundreds of academic researchers worldwide.

- Led two parallel empirical programs establishing Reddit's community quality measurement: validated survey instruments and hierarchical models for belonging, values alignment, and governance across **2.8K users and 281 subreddits** (AAAI 2025), establishing a company-wide "successful communities" top-line metric; and mixed-methods research on **1K subreddit founders** (CHI 2024) that precipitated restructuring of a 60-person Community Org and a **\$5M+ annual investment** in community formation products.
- Led a novel within-community, pre-registered RCT across 98K users and 33 subreddits to measure the impact of proactive, community-specific posting guidance; found meaningful gains in submission success and downstream engagement, with substantially fewer AutoModerator removals and no reduction in participation.

Twitch / Amazon · San Francisco, CA · 2016–2022

Head of Science, Community Health · Senior Research Scientist, Central Science

Built and led a 5-person Research and Data Science team, Community Health's first dedicated science function, covering moderation measurement, safety ML evaluation, and user-reporting systems at platform scale; reported to the VP of Community Health.

- Designed a **harm-prevalence measurement pipeline** combining model-assisted sampling with expert labeling workflows to estimate policy-violation rates across billions of daily chat messages, including severe harm categories with **base rates as low as 1 in 10,000**; became the org's primary safety measurement instrument, created gold-standard datasets for moderation ML training and evaluation, and informed staffing decisions during a company-wide RIF.

- Built a multi-year research program on **moderator workflows, expertise, and needs**, spanning qualitative studies of decision-making and reflective practice (CSCW 2022), quantitative validation via a representative survey of 1,053 moderators stratified by channel size (GROUP 2023), and survey/log-data studies to inform platform-wide moderation-tooling strategy.
- Developed a **community health measurement framework** integrating both harm signals (the prevalence pipeline above) and positive signals: validated sense-of-community instruments combining survey and behavioral trace data across ~1,400 viewers and 295 channels (CSCW 2022); hierarchical model predicted one-year viewer retention and surfaced design implications for belonging at platform scale.
- Founded and led the **Twitch Fellowship**, a grant program awarding \$10K research fellowships to PhD students studying online communities, creator economies, and platform health — building an external research ecosystem and positioning Twitch as a partner for academic inquiry.

EDUCATION

Stanford University · Stanford, CA

PhD, Computer Science — Human-Computer Interaction & Social Computing

Dissertation: "Understanding and Supporting Selective Sharing" · Advisor: Jeffrey Heer

MA, Philosophy · BS, Mathematics; Minor, Symbolic Systems

EARLIER RESEARCH EXPERIENCE

Microsoft Research · Research Intern · Redmond, WA · 2012, 2013 · *Meredith R. Morris, Jaime Teevan, Susan Dumais*

Yahoo Labs · Academic Contractor · San Francisco, CA · 2015–2016 · *David A. Shamma, Jofish Kaye*

Palo Alto Research Center (PARC) · Research Assistant · Palo Alto, CA · 2008–2010 · *Peter Pirolli, Ed H. Chi*

Facebook · Research Intern, Core Data Science · Menlo Park, CA · 2013 · *Dan Merl*

Google · Research Intern, User Experience Research · Mountain View, CA · 2011 · *Ed Chi, Mike Brzozowski*

SELECTED PUBLICATIONS

- Horta Ribeiro, M., West, R., Lewis, R., & **Kairam, S.** (CSCW 2025): Post Guidance for Online Communities. — *Conducted the first large-scale RCT showing that proactive, community-specific guidance at contribution time shifts compliance from enforcement to prevention — improving content quality, reducing moderator burden, and increasing post success rates without reducing participation.*
- Weld, G., Pearson, C., Spahn, B., Althoff, T., Zhang, A.X., & **Kairam, S.** (AAAI 2025): How Conversational Structure and Style Shape Online Community Experiences. — *Developed a content-agnostic behavioral framework for predicting community belonging across diverse online contexts, showing that linguistic style alone explains >30% of variance — enabling scalable evaluation of how AI-mediated interactions affect social cohesion.*
- Cunningham, T., Pandey, S., Sigerson, L., Stray, J., Allen, J., Barrilleaux, B., Iyer, R., Kothari, M., Rezaei, B., & **Kairam, S.** (Annals NYAS, 2025): Ranking by Engagement and Non-Engagement Signals: Learnings from Industry. — *Unique cross-industry partnership demonstrating how optimizing for short-term engagement leads recommendation systems and platforms astray — and how incorporating non-engagement signals can counteract these effects, with direct implications for how AI systems shape societal outcomes.*
- **Kairam, S.** & Foote, J. (CHI 2024): How Founder Motivations, Goals, and Actions Influence Early Trajectories of Online Communities. — *Revealed that founder motivations and early governance choices propagate durably through community trajectories — providing an empirical framework for anticipating how founding-moment design decisions shape long-term AI deployment dynamics.*
- Seering, J. & **Kairam, S.** (GROUP 2023): Who Moderates on Twitch and What Do They Do? — *First representative large-scale quantification of who moderates online communities and how — overturning assumptions about moderation labor and surfacing systematic gaps in AI-assisted governance tooling design.*
- Cullen, A. & **Kairam, S.** (CSCW 2022): Practicing Moderation: Community Moderation as Reflective Practice. — *Showed that effective governance requires reflective expertise — contextual judgment, learning-by-doing, and collective sensemaking — that is systematically degraded when moderation is automated without preserving the capacity for human oversight and deliberation.*
- **Kairam, S.**, Mercado, M.C., & Sumner, S.A. (CSCW 2022): A Social-Ecological Approach to Modeling Sense of Virtual Community in Livestreaming Communities. — *CDC collaboration initiating cross-sector partnership to validate multi-dimensional community health instruments; combining survey and behavioral trace data across ~1,400 viewers and 295 channels to predict one-year viewer retention and measure positive social outcomes at scale.*
- **Kairam, S.** & Heer, J. (CSCW 2016): Parting Crowds: Characterizing Divergent Interpretations in Crowdsourced Annotation Tasks. — *Developed a novel computational method to surface and characterize systematic annotator disagreement — reframing disagreement as structured, meaningful signal rather than noise, with direct implications for LLM evaluation design, evaluator calibration, and preference data collection.*

SKILLS

Methods: experimentation & A/B testing, Bayesian & mixed-effects modeling, annotation quality & evaluation rubric design, behavioral trace analysis, causal inference, survey design & analysis, prevalence estimation & model-assisted sampling, mixed-methods research

Languages & Tools: Python, SQL, R

RECENT INVITED TALKS & PANELS

Recent invited talks on AI, online governance, and sociotechnical systems.

Cornell Tech: LLM & Society Workshop (Participant, May 2025) · **Princeton CIP Seminar** (Invited Lecture, Apr 2025) · **University of Washington INFO 492A** (Invited Lecture, Apr 2025) · **CSCW 2024** (Social Computing SIG Organizer, Nov 2024) · **CHI 2024** (Paper Presentation, May 2024) · **Princeton COS 436** (Invited Lecture, Sep 2024) · **IC2S2 2023** (Paper Presentation, Jul 2023)