Jackson Phillip Trager

Google Scholar | LinkedIn | Email | Jacksonptrager.com

Research Topics

AI Ethics ◆ LLM Resources & Evaluation ◆ HCI ◆ Threat ◆ Cultural Morality ◆ Conflict ◆ Fairness ◆ Applied AI

Research Methods and Skills

Behavioral Experiments ◆ Annotated Corpora Curation ◆ AI Error Analysis & Evaluations ◆ Social Media Analysis
Survey Design ◆ R Programming ◆ Data Analysis ◆ Anthropological Fieldwork ◆ Qualitative Interviewing

Research Experience

2025 **Summer Research Associate - AI, Culture, & Government,** RAND Corporation, Santa Monica

- Analyzed perceptions and impacts of emerging technologies and AI across communities and cultures for the Science and Technology Dept and the National Security Research Division

2020- **Behavioral Researcher,** Morality and Language Lab - Computational Social Science, *University of Southern California*

- Led 5+ published research projects employing mixed methods—including behavioral experiments, survey analysis, and NLP computational methods—to explore culture, conflict, and technology within and across cultures
- Managed teams of 20+ annotators across multiple projects, developing diverse LLM training corpora with over 40,000 annotated posts in multiple languages, focusing on subjective measurements of morality, hate, and information spread

2021- **Human-Centered AI Researcher for Sociotechnical Systems,** Everyday Respect LAPD Project, *USC*

- Directed the human-centered AI team for the Everyday Respect Project, building multi-perspective AI models that analyze LAPD body cam data to support transparency and accountability in sociotechnical system improvements with bias mitigation

2024 **LLM Integration - User Experience Researcher,** AgentGrow, Los Angeles

- Led UX research leveraging user feedback surveys, behavioral metrics, LLM fine-tuning strategies, and prompt engineering to enhance product design, address ethical concerns, and improve trust and safety

2022 **Annotation Project Manager,** Google Jigsaw + USC

- Led a team of annotators to label 16,099 comments for hate-based rhetoric to utilize for LLM fine-tuning and evaluation

2016-18 **Applied Anthropology Researcher,** Institute of Cognition & Culture, *Queen's University Belfast, U.K.*

- Mixed-methods research on shifting intergroup relations and cross-community outreach of the Northern Irish conflict

2015 **Psychology Research Assistant,** The Herczeg Institute, *Tel Aviv University, Israel*

- Data analysis of hostile-world scenarios, well-being, and narrative/sentiment on the collective trauma of Israelis/Palestinians

Selected Publications

AI Alignment

"The Shrinking Landscape of Linguistic Diversity in the Age of Large Language Models" (2025) (R&R, Nature Human Behavior)
"Perils and Opportunities in Using LLMs in Psychological Research" (2024) PNAS NEXUS
"Improving Communication in the Shadow of Power: New (AI) Technologies & Policing" (2023) (under review, Nature)
"Teaching Humans to Teach AI: A Human-Centered Framework for Responsible Subjective AI in Sociotechnical Systems" (in prep)

LLM Evaluation

"MFTCXplain: A Multilingual Benchmark Dataset for Evaluating the Moral Reasoning of LLMs via Hate Speech Multi-hop Explanations " (2025) Findings of the Association for Computational Linguistics: EMNLP
"Semantic F1 Scores: Fair Evaluation Under Fuzzy Class Boundaries" (2025) (under review, ICLR)
"The Moral Foundations Reddit Corpus" (2022) arXiv
"Generative Agents and Intergroup Conflict" (2025)(in prep)

Morality Across Culture & Technology

"Moral values predict county-level COVID-19 vaccination rates in the US" (2022) American Psychologist
"The Immorality of Excessive Wealth Across Culture" (2025) PNAS NEXUS
"Moral Alignment Shapes Responses to Shared Content" (2023) Journal of Experimental Psychology
"The (Moral) Language of Hate" (2023) PNAS NEXUS
"Hate Is Justified When Values Are Threatened: Evidence from Threatening Tweets and Real World Events" (2024) arXiv

RAND Policy Reports

"Using AI to Understand Public Perceptions: How LLMs Could Help the USG Anticipate Views of Emerging Technologies"
"A Social Science Methods Toolkit for the U.S.G. to Gauge Public Perception of Technologies Across Development Cycle"
"Beyond Borders: Improving Personnel Vetting for International Applicants"

Consulting & Teaching Experience

2020-22 **Psychology Lecturer,** *University of Southern California* – Los Angeles

- Psychological Statistics, Experimental Psychological Research Methods, Psych Science and Society, Cognitive Processes

2018-21 **Anthropology Lecturer,** *Pierce College, Harbor City College, Valley College* – Los Angeles

- Cultural Anthropology, Anthropology of Religion, Human Systems

2019-21 **Director of Curriculum,** The Antedote Initiative - A Wholeness Program for Adolescence

- Led the development of a curriculum for a program addressing the complex emotional, psychological, and behavioral challenges in adolescence, focusing on managing anxiety, depression, technology, and minor substance abuse

2019-21 **Applied Culture Development Consultant,** WolfTribe - Los Angeles, California

- Consulted on corporate workplace culture and team development for Sony, Dreamworks, USC Marshall Business School, Microsoft, Redbull, SpaceX, Snapchat, Walt Disney, Mattel, NorthropGunman, Google, Statefarm

2013-19 **Culture Development Consultant,** Fulcrum Adventures

- Led workshops on workplace culture, conflict resolution, and strategy implementation for LAUSD, UCLA, CSLA, USC, Apple, Universal, Medtronic, Fox, Farmers Insurance, Wounded Warriors, Paramount, Virgin Galactica

2015-16 **Interfaith Professional Fellow,** Newground's Muslim-Jewish Fellowship for Change, Los Angeles

- Collaborated with the White House Advisory Council on Faith-Based & Neighborhood Partnership to combat digital hate

Education

2025/6 PhD Social Psychology (expected), *University of Southern California*
2023 MA Social Psychology, *University of Southern California*
2017 MA Cognitive Anthropology, *Queen's University Belfast*, Northern Ireland, U.K.
2015 BA Psychology & Religious Studies, *CSU Northridge*, Cum Laude - Honors - Minor: Cultural Studies

International Field Experience

LAPD Tech/Constitutional Policing (2023-25); Manila, Philippines (2019); Belfast, Northern Ireland (2016-17); Malaysia-Myanmar (2016); Israel - Palestine (2015); Armenia (2015); Poland-Lithuania (2014)

Grants, Honors, & Scholarships

2025 Plurality Institute Grant Award for LLMS Improving Democratic Discourse, Technology Race and Prejudice Awardee
2024 Center for Computational Linguistics Studies Finalist, Psychology Department Doctoral Research Grant Award
2023 USC Dornsife Dean's Emblem Award for Outstanding Scholarship
2022 Psychology Department Doctoral Research Grant Award (4 years), Graduate Student Research Fellowship USC (4 years)
2018 Distinction in Graduate Research Dissertation Queen's University Belfast
2016 LA Human Relations Commission Certificate of Community Engagement, International Postgrad Scholarship
2015 Newground's Muslim-Jewish Fellowship, Onward Israel JFed Grant, Religious Studies Scholar Award,
CS Honors Conference Scholarship, AS Medallion, Mellon Foundation Representative
2014 AAR Religion Competition Award, US Society of Leadership & Success, JS Scholarship, IRA Scholarship, Crisis Helpline Cert.
2010 Cal Grant Awardee (4 years)
2009 Eagle Scout Boy Scouts of America