Proposal: ICLR 2025 Workshop on Human-AI Coevolution

Jacy Reese Anthis[1,2], Dylan Asmar[1], Katie Driggs-Campbell[3], Amelia Francesca Hardy[1], Kiana Jafari Meimandi[1], Geoff Keeling[4], Mykel Kochenderfer[1], Houjun Liu[1], Shuijing Liu[5], Roberto Martin-Martin[5], Ahmad Rushdi[1], Marc Schlichting[1], Peter Stone[5], Hari Subramonyam[1], and Diyi Yang[1]

[1]Stanford University
[2]University of Chicago
[3]University of Illinois Urbana-Champaign
[4]Google
[5]The University of Texas at Austin

1 Workshop Description

This workshop aims to build a multidisciplinary research community around the emerging field of human-AI coevolution (HAIC) to understand the feedback loops that emerge from continuous and long-term human-AI interaction. As AI systems have become more prevalent and have been present in society over longer periods, scholars from diverse fields and methodologies have come to focus on HAIC and its importance for system architecture, human feedback, regulation, and other domains (e.g., Damiano & Dumouchel, 2018; Järvelä et al., 2023; Matsubara et al., 2023; Donati, 2021; Zhao et al., 2024). Through this workshop we hope to lay a collaborative foundation for this research agenda. To achieve this we will organize expert talks from academia and industry, dynamic panel discussions, interactive breakout sessions, and networking opportunities, drawing on our diverse experience organizing related workshops at leading conferences in ML, NLP, HCI, and related fields.

Across diverse domains, including algorithms, recommendation systems, and large language models (LLMs), this workshop challenges the traditional view of AI solely as a tool that improves through human-provided signals (Anthis et al., 2024; Chang et al., 2024; Kulkarni & Rodd, 2020; Mehrabi et al., 2021; Chang et al., 2023; Meimandi et al., 2023); instead, it will also investigate how humans also alter their behaviors, decision-making processes, and cognitive frameworks in response to long-term interactions with AI and how AI systems can be developed in response to changes in human feedback over time (Gabriel et al., 2024; Subramonyam et al., 2024; Wu et al., 2023; Zhao et al., 2024).

Research on HAIC entails a move beyond typical performance metrics of AI benchmarks, exploring multiple levels of analysis. From a low-level perspective, HAIC can occur as a single human and AI agent interact over time as bidirectional learning processes reshape behavior (Liu et al., 2024a; Maples et al., 2024; Mozannar et al., 2023; Reuel et al., 2024b). This co-evolution also appears at a modeling level: as "gold" web-scale training datasets becomes contaminated with already-AI generated output, new behaviors and risks emerge (Gerstgrasser et al., 2024; Shumailov et al., 2024). From a high-level perspective, it can involve long-term interaction across many humans (Ge et al., 2024; Liu et al., 2024c) and AI agents (Park et al., 2023; Wu et al., 2023) and its impact on social institutions, such as on healthcare (Bica et al., 2021; Grote & Keeling, 2022; Vaidyam et al., 2019), education (Roll & Wylie, 2016; Yang et al., 2013, 2015), transportation (Keeling et al., 2019; Keeling, 2020; Liu et al., 2022, 2023), and criminal justice (Jacobs & Wallach, 2021; Marx et al., 2020). This multidisciplinary, multilevel approach is reflected in the research questions, themes, and selection of the expert panelists and speakers. Open questions we will discuss and debate include:

Human-AI Interaction and Alignment
How do human expectations, reliance, and trust in AI systems evolve as they interact with machine learning models in real-world contexts such as healthcare or finance? How can systems be designed to

align with human intentions and values, ensuring that these systems are trustworthy and foster positive human-AI interactions over time and across cultures (Ge et al., 2024; Sorensen et al., 2024)? What are the ethical and societal implications of HAIC? How does HAIC affect human autonomy (Hong & Williams, 2019) and social norms (Boman, 1999), and what are the implications for AI developers who aim to improve future human-AI interactions?

Algorithmic Adaptation and Robustness
How can methods like Reinforcement Learning from Human Feedback (RLHF) be improved and expanded to adapt to changing preferences (Carroll et al., 2024), avoiding sycophancy (Kirk et al., 2023), and other complexities of interaction? What technical methods can ensure that these systems adapt in a way that avoids reinforcing biases while promoting intentional and thoughtful decision-making? How can AI systems be made robust—that is, capable of maintaining performance across diverse and unforeseen contexts and tasks?

Long-Term Societal Impact and Safety
What are the broader implications of HAIC on socio-technological systems, including governance (Reuel et al., 2024a), policy, and collective decision-making Liu et al. (2022, 2023, 2024b)? How can AI alignment principles be incorporated into these systems to ensure that they remain beneficial to society and aligned with human values over time? How do traditional approaches to AI safety need to be rethought in light of the dynamic interplay between human and AI systems? What is the impact of existing AI systems, and what does this tell us about future potential? What role can AI systems play in forecasting and mitigating possible negative consequences, thereby enhancing safety in critical sectors?

2 Key Themes

2.1 Bidirectional Learning Beyond Performance Metrics

Rather than focusing solely on improving AI systems in terms of isolated environments or benchmarks, this workshop will explore how human cognition, decision-making, and behavior evolve through prolonged interaction with AI systems. This includes how people adjust their trust, how AI exacerbates and mitigates human bias, and how reliance on AI varies across different settings (e.g., healthcare, law enforcement, personal decision-making). As AI systems become agentic (Chan et al., 2023; Kenton et al., 2022; Meimandi et al., 2024) and capable of longitudinal interactions, we also investigate their influence can go beyond shaping individual behavior and reconfigure socio-technological norms.

An essential aspect of our exploration is understanding how to revise or redesign AI systems and algorithms to adopt a bidirectional learning approach, where not only do humans learn from AI, but AI systems also continuously adapt to and learn from human feedback, behavior, and values—leading to profound changes in how humans relate to technology and to one another. This includes the design of experiments and metrics to evaluate AI systems in light of HAIC, in particular, to avoid critical issues of AI learning negative patterns, such as Microsoft's Tay chatbot that parrotted hate speech and other negative content when released on Twitter in 2016 (Lee, 2016).

2.2 Shaping Collective Behavior and Learning

Many applications of modern AI systems, such as in classrooms, workplaces, and democratic deliberation, occur in the presence of multiple humans and often include various AI models. In contexts such as group collaboration (e.g., remote teamwork), policy-making, or public interaction with AI systems, the feedback loops between AI and human teams could shape how decisions are made, how consensus is built, and how biases are formed or mitigated over time. For example, recent work by Tessler et al. (2024) developed a "Habermas machine," in which LLMs summarize diverse viewpoints in a manner that attempts to balance majority and minority viewpoints. The system then allows the audience to critique its synthesis, iteratively fashioning collective expressions and potentially reshaping political conversations and collective decision-making. In light of the difficulty of creating explainable models, understanding the subtle and implicit influences of AI systems during such interactions will pose a unique and important challenge.

2.3 Dynamic Feedback Loops in Socially Impactful Domains

This workshop will investigate how real-time feedback loops in high-stakes interactions between humans and AI shape task outcomes, agent behaviors, and even societal structures. This includes decisions (e.g., hiring, medical diagnosis, recidivism prediction) that affect long-term outcomes for humans as well as everyday collaborative interactions. In domains such as education, systems have emerged that emulate student behavior for instructors Wang & Demszky (2023) or directly provide instruction, such as Khanmigo, built by Khan Academy with GPT-4 (OpenAI, 2023). These broad applications and use cases entail unique, domain-specific demands and considerations, such as patient safety in healthcare, fairness in legal systems, mental burden in human-robot collaboration, and personalization in education. Addressing these needs requires deliberately tailored AI-human interactions that will shape the distinct ways AI systems and humans co-evolve in each context.

2.4 Socio-Technological Bias, Norms, and Ethics

AI systems are inherently influenced by training data and algorithmic design, often perpetuating biases and reinforcing specific societal norms, such as the differences in automatic speech recognition across racial groups (Koenecke et al., 2020) or the heavier moderation of content created by minoritized groups (Binns et al., 2017; Dias et al., 2021; Haimson et al., 2021; Liu et al., 2024b). This workshop will examine how AI systems can effect significant representational harm (Dastin, 2018; Hao, 2019; Mengesha et al., 2021) but also clarify and mitigate human bias (Celiktutan et al., 2024; Kleinberg et al., 2020). The human norms on which society operates, such as what is considered polite, respectful, or embarrassing (Reeves & Nass, 1996), may be reshaped, such as when job applicants and prospective students are able to send out personalized emails en masse. Ethical concerns will be discussed, particularly regarding how feedback loops between humans and AI can either improve fairness and transparency or entrench harmful patterns, amplifying existing biases when deployed in critical decisions such as college admissions (Alvero et al., 2020). Concepts such as "bias," which are already well-established as being often oversimplified in ML research (Blodgett et al., 2020), will become more complex—raising pressing and time-sensitive questions for researchers and practitioners.

3 Workshop Format

Invited Talks: These will be presentations by leading experts invited to discuss key topics related to the theme of HAIC. Each invited talk will last for 30 minutes, including 25 minutes for the presentation and 5 minutes for Q&A. The aim is to provide deep insights from researchers and industry professionals at the forefront of the field, setting the tone for the workshop.
Spotlight Talks: Spotlight talks are shorter, 15-minute presentations selected from submissions. They are designed to highlight emerging research relevant to HAIC. These talks provide an opportunity for more participants to share their work, encouraging a diverse range of perspectives and innovative ideas.
Panel Discussions: There will be two 1-hour panel discussions, where experts from diverse backgrounds will explore the challenges, opportunities, and the future of HAIC. The panels will foster a dynamic discussion involving both the panelists and the audience, focusing on complex issues like bidirectional learning, collective behavior and institutions, and safety and ethics in light of HAIC.
Breakout Groups: Breakout sessions are interactive segments aimed at encouraging deeper discussion in smaller, more focused groups. These sessions will enable participants to engage in collaborative activities, share experiences, and work together on practical challenges. This format ensures that attendees have an opportunity to network, dig deeper into specific questions, and contribute more actively to discussions

4 Diversity Commitment

We are committed to promoting diversity and inclusion in all aspects of our workshop. In selecting organizers and speakers, we have prioritized diversity across gender, race, institutional affiliation, and scientific discipline. Our roster includes individuals from various backgrounds and career stages, ensuring representation from undergraduate, masters, Ph.D. candidates, postdoctoral researchers, and

industry researchers, as well as assistant, associate, and full professors. This diverse group brings a wide range of perspectives and experiences, enriching the workshop's overall quality.

Our submission review process will be double-blind, conducted through OpenReview, to minimize institutional and author biases. The program will be curated to feature a wide representation of research topics while adhering to high standards of quality. This approach will help us assemble a diverse group of participants, with selected contributors given the opportunity to present alongside our primary invitees.

Our workshop will include a Tiny Papers track to attract and support underrepresented, under-resourced, and early-career researchers. Consistent with ICLR's 2024 Tiny Papers initiative, we will require that every Tiny Paper submission has at least one key author who self-identifies as an under-represented minority (URM). To encourage submissions, we will publicize our workshop and this track within affinity groups at our institutions and their partner groups at other institutions worldwide. Additionally, our web page will answer Tiny Paper FAQs and provide a directory of research mentorship resources for underrepresented minority scholars.

We plan for the workshop to be an in-person event, but to account for extenuating circumstances as well as engagement before and after the workshop, we will incorporate hybrid synchronous and asynchronous elements to maximize participation. Invited speakers and authors of accepted papers will be encouraged to provide pre-recorded presentations to allow flexible access for attendees. Live components, including panel discussions and Q&A sessions, will be facilitated using platforms like sli.do to encourage interactive engagement.

To further expand the accessibility of our workshop, we plan to offer registration fees and travel grants to individuals who face financial barriers to attending. To support this initiative, we will seek sponsorships from leading organizations with whom we have existing collaborations and/or funding relationships, including Google DeepMind, Meta AI, Accenture, LVMH, Microsoft, NASA, Airbus, and Qualcomm.

Additionally, we will establish a mentorship program that pairs early-career researchers with senior academics before or during the workshop. This initiative aims to foster long-term professional relationships and ensure that early-stage researchers feel supported and included in the community.

We will also develop and communicate a clear code of conduct to create a welcoming and respectful environment for all participants. Our goal is to establish shared expectations for behavior and promote a positive and inclusive atmosphere throughout the workshop.

5 Invited Speakers and Panelists

Our workshop will feature invited talks and panels from a diverse group of researchers, representing various backgrounds, institutions, and specialties. These speakers and panelists will offer distinct viewpoints on the latest advancements in HAIC and associated subjects. We intend to announce the titles of all presentations and agendas for panel discussions before the event begins. We have tentatively confirmed the following speakers.

- **AJ Alvero (Cornell University)**: computational sociology, linguistic sociology, education, culture and bias
- **Patrick Connolly (Accenture)**: AI governance, global responsible AI, generative AI research
- **James Evans (University of Chicago / Santa Fe Institute / Google)**: computational social science, collective intelligence, machine learning and AI, human-machine intelligence, human and machine languages
- **Sara Hooker (Cohere AI)**: language modeling, interpretability, learning from human feedback, alignment, fairness, toxicity mitigation
- **Dorsa Sadigh (Stanford University)**: efficient algorithms for safe, reliable, adaptive human-robot and generally multi-agent interaction
- **Lindsay Sanneman (Massachusetts Institute of Technology)**: artificial intelligence, robotics, human-robot interaction, human factors

- **Sarah Sebo (University of Chicago)**: social dynamics in human-robot and human-computer interaction
- **Abigail See (DeepMind)**: natural language processing, controllability, interpretablility, coherence of neural text generation
- **Cong Shen (University of Virginia)**: machine learning, in-context, distributed and decentralized learning, wireless communications, networking
- **Fan Shi (National University of Singapore)**: robotics, AI safety
- **Winnie Street (Google)**: artificial intelligence, cognitive science, psychology

6 Tentative Schedule

This workshop will follow a hybrid format. It will feature four invited talks, each lasting 30 minutes (with 25 minutes for presentation and 5 minutes for Q&A), as well as four 15-minute spotlight talks selected from submissions. There will also be two 1-hour poster sessions and two 1-hour panel discussions exploring the challenges, opportunities, and future of human-AI coevolution. This will primarily be an in-person event, but we will have a remote option for extenuating circumstances, such as last-minute travel interruptions and illness. We have an anticipated audience size of 100, judging from outreach we have done so far to promote the workshop and will continue to do (e.g., social media posts, Slack posts, direct contact).

The tentative workshop schedule is stated in Table 1. We will be following the suggested timeline by ICLR:

- Submission Date for Workshop Contribution: 3 February 2025
- Accepted Paper Notification Date: 5 March 2025, 11:59 PM AoE
- Final Workshop Program, Camera-Ready, Video Uploads: 27 March, 11:59 PM AoE

Morning Session		**Afternoon Session**	
08:50 - 09:00	Welcome Notes	13:00 - 13:30	Invited Talk 3
09:00 - 09:30	Invited Talk 1	13:30 - 14:00	Invited Talk 4
09:30 - 10:00	Invited Talk 2	14:00 - 14:15	Spotlight Talk 3
10:00 - 10:15	Spotlight Talk 1	14:15 - 14:30	Spotlight Talk 4
10:15 - 10:30	Spotlight Talk 2	14:30 - 15:15	Break + Poster Session 2
10:30 - 11:15	Break + Poster Session 1	15:15 - 16:15	Panel 2
11:15 - 12:15	Panel 1	16:15 - 17:15	Breakout Sessions
12:15 - 13:00	Lunch break	17:15 - 17:30	Closing remarks

Table 1: The Tentative Workshop Schedule

7 Previous Related Workshops

1. 2nd Workshop on Generative AI and Law (GenLaw '24) (ICML 2024)
2. Agent Learning in Open-Endedness Workshop (NeurIPS 2023)
3. AI for Agent-Based Modelling (ICLR 2023)
4. AI meets Moral Philosophy and Moral Psychology: An Interdisciplinary Dialogue about Computational Ethics (NeurIPS 2023)
5. Algorithmic Fairness through the Lens of Time (NeurIPS 2023)
6. Aligning Reinforcement Learning Experimentalists and Theorists (ICML 2024)
7. Artificial Intelligence & Human Computer Interaction (ICML 2023)
8. Challenges in Deployable Generative AI (ICML 2023)
9. Foundation Models for Decision Making (NeurIPS 2023)
10. From Cells to Societies: Collective Learning Across Scales (ICLR 2022)

11. Humans, Algorithmic Decision-Making and Society: Modeling Interactions and Impact (ICML 2024)
12. ICML 2024 Workshop on Foundation Models in the Wild (ICML 2024)
13. Instruction Tuning and Instruction Following (NeurIPS 2023)
14. Models of Human Feedback for AI Alignment (ICML 2024)
15. Socially Responsible Language Modelling Research (SoLaR) (NeurIPS 2023)
16. Socially Responsible Machine Learning (ICLR 2022)
17. The 4th Workshop on practical ML for Developing Countries: learning under limited/low resource settings (ICLR 2023)
18. The Many Facets of Preference-Based Learning (ICML 2023)
19. Trustworthy and Reliable Large-Scale Machine Learning Models (ICLR 2023)
20. Trustworthy Multi-modal Foundation Models and AI Agents (TiFA) (ICML 2024)
21. Workshop on Large Language Models for Agents (ICLR 2024)
22. Workshop on Theory of Mind in Communicating Agents (ICML 2023)

8 Organizers and Biographies

Jacy Reese Anthis (Stanford University / University of Chicago)

- **Email:** anthis@stanford.edu
- **Webpage:** https://jacyanthis.com
- **Google Scholar:** https://scholar.google.com/citations?user=lRhxKSAAAAAJ
- **Bio:** Jacy Reese Anthis is the director of the Sentience Institute, a visiting scholar at the Stanford Institute for Human-Centered Artificial Intelligence (HAI), and a PhD candidate in the sociology and statistics departments at the University of Chicago. His research, published in venues such as CHI, HRI, and NeurIPS, lies at the intersection of human-computer interaction and machine learning with a particular focus on the design and evaluation of human-like AI systems that appear to have mental faculties such as reasoning, emotion, and agency. He is active in HCI and ML conferences, including most recently as a reviewer for ICLR 2025, an Area Chair for CSCW 2025, and co-organizing a CSCW 2024 workshop on AI red teaming (Zhang et al., 2024) that drew on his academic research and contract role as a red teamer for OpenAI.

Dylan Asmar (Stanford University)

- **Email:** asmar@stanford.edu
- **Webpage:** https://www.linkedin.com/in/dylanasmar/
- **Google Scholar:** https://scholar.google.com/citations?user=3XO3Jb4AAAAJ
- **Bio:** Dylan Asmar is a Hugh H. Skilling Stanford Graduate Fellow and Ph.D. candidate in Aeronautics and Astronautics at Stanford University, working in the Stanford Intelligent Systems Laboratory (SISL). He holds an M.S. in Aeronautics and Astronautics from MIT. His research focuses on decision making under uncertainty, emphasizing human-AI collaboration and multiagent systems. Drawing on his experience evaluating technologies in aviation, Dylan develops decision-making frameworks that aim to optimize performance and reliability by combining the strengths of algorithms and human expertise.

Katie Driggs-Cambell (University of Illinois Urbana-Champaign)

- **Email:** krdc@illinois.edu
- **Webpage:** https://krdc.web.illinois.edu/
- **Google Scholar:** https://scholar.google.com/citations?user=UXNLsZUAAAAJ
- **Bio:** Katie Driggs-Campbell received her BSE in Electrical Engineering from Arizona State University and her MS and PhD in Electrical Engineering and Computer Science from the University of California, Berkeley. She is currently an Assistant Professor in the ECE Department at the University of Illinois at Urbana-Champaign. Her research focuses on exploring and uncovering structure in complex human-robot systems to create more intelligent, interactive autonomy. She draws from the fields of optimization, learning & AI, and control theory, applied to human-robot interaction and autonomous vehicles.

Amelia Francesca Hardy (Stanford University)

- **Email:** ahardy@stanford.edu
- **Webpage:** https://www.linkedin.com/in/ameliahardy/
- **Google Scholar:** https://scholar.google.com/citations?user=YS-7RFUAAAAJ
- **Bio:** Amelia Hardy is a Master's student in Computer Science at Stanford University, affiliated with the Stanford Intelligent Systems Laboratory (SISL), Center for Research in Foundation Models (CRFM), and the Stanford NLP group. Her research focuses on language model evaluation and reinforcement learning. She has led research projects in these areas, most recently in collaboration between SISL and Airbus for an investigation of LLM's and aviation. She has served as a reviewer for the Journal of Aerospace Information Systems (JAIS) and ACM Intelligent User Interfaces (IUI). She was a member of the second-place Amazon Alexa Prize team. Previously, Amelia earned her BS in Computer Science at Stanford and worked as an ML Engineer at Woebot, a startup co-founded by Andrew Ng.

Kiana Jafari Meimandi (Stanford University)

- **Email:** kjafari@stanford.edu
- **Webpage:** https://www.linkedin.com/in/kiana-jafari/
- **Google Scholar:** https://scholar.google.com/citations?user=R_BnJ3YAAAAJ
- **Bio:** Kiana Jafari Meimandi is a postdoctoral researcher at the Stanford Intelligent Systems Laboratory (SISL). She has a master's degree in industrial engineering from KNTU, a master's degree and Ph.D. in systems engineering from the University of Virginia. Her research focuses on human-agent teaming and specifically how to incorporate socio-technological norms into AI agent's learning algorithms. Her research interests also include designing responsible, and robust AI systems. She has been on the Reviewer Board of MDPI (Multidisciplinary Digital Publishing Institute), Machine Learning and Knowledge Extraction since 2019. She served as the session chair and poster judge for SIEDS. Kiana also worked as the session chair, and robotics competition coordinator at IEEE RO-MAN 2024. She served as a reviewer for more than 10 journals and conferences including ACM JATS, IEEE THMS, and IEEE ICHMS.

Geoff Keeling (Google)

- **Email:** gkeeling@google.com
- **Webpage:** https://geoffkeeling.github.io
- **Google Scholar:** https://scholar.google.com/citations?user=_k8b6mYAAAAJ
- **Bio:** Geoff Keeling is a Senior Research Scientist at Google and an Associate Fellow at the Leverhulme Centre for the Future of Intelligence at the University of Cambridge. His research explores the ethics and cognitive science of frontier AI systems. Prior to Google, Geoff was a Research Fellow at Stanford University, based between the Institute for Human-Centered AI and the McCoy Family Center for Ethics in Society. He has organized a variety of events and workshops related to AI assistants, computer ethics, and affective computing.

Mykel Kochenderfer (Stanford University)

- **Email:** mykel@stanford.edu
- **Webpage:** https://mykel.kochenderfer.com/
- **Google Scholar:** https://scholar.google.com/citations?user=cAy9G6oAAAAJ
- **Bio:** Mykel Kochenderfer is Associate Professor of Aeronautics and Astronautics and Associate Professor, by courtesy, of Computer Science at Stanford University. He is the director of the Stanford Intelligent Systems Laboratory (SISL), conducting research on advanced algorithms and analytical methods for the design of robust decision making systems Prior to joining the faculty in 2013, he was at MIT Lincoln Laboratory where he worked on airspace modeling and aircraft collision avoidance. He is affiliated with the Stanford Artificial Intelligence Laboratory (SAIL), the Human-Centered AI (HAI) Institute, the Symbolic Systems Program, the Bio-X Institute, Wu Tsai Neurosciences Institute, and the Center for Automotive Research at Stanford (CARS). In 2017, he was awarded the DARPA Young Faculty Award. In 2023, he served as General Chair for the Learning for Dynamics and Control Conference. He is chair of the editorial board of the Journal of Artificial Intelligence Research and an associate editor of the Journal of Aerospace Information Systems.

Houjun Liu (Stanford University)

- **Email:** houjun@stanford.edu
- **Webpage:** https://nlp.stanford.edu/~houjun/
- **Google Scholar:** https://scholar.google.com/citations?user=YS-7RFUAAAAJ
- **Bio:** Houjun Liu is an undergraduate student in computer science at Stanford University, affiliated with the Stanford NLP Group and Stanford Intelligent Systems Lab (SISL). He is also a NIH-funded consulting research software engineer at CMU TalkBank corpus for multilingual speech and language pathology analysis. His research focuses on both discovering mechanistically-informed optimizations of language modeling and sequential decision formulations of natural language understanding. He has served as a reviewer for Alzheimer's Research and Therapy, Journal of Aerospace Information Systems (JAIS), and on the program committees of workshops for NeurIPS and AAAI.

Shuijing Liu (The University of Texas at Austin)

- **Email:** shuijing.liu@utexas.edu
- **Webpage:** https://shuijing725.github.io/
- **Google Scholar:** https://scholar.google.com/citations?user=I4k7ukgAAAAJ
- **Bio:** Shuijing Liu is a postdoctoral scholar in the Computer Science Department at The University of Texas at Austin. She received her Ph.D. in Electrical and Computer Engineering Department at University of Illinois Urbana-Champaign in 2024. Her interest is at the intersection of human-centered robotics and machine learning. Specifically, Shuijing works on learning interaction models for robot navigation in challenging human environments. Besides, she is also interested in human-robot interaction through language. Prior to her Ph.D., she earned a Bachelor's degree in Computer Engineering at University of Illinois Urbana-Champaign.

Roberto Martin-Martin (The University of Texas at Austin)

- **Email:** robertom@cs.utexas.edu
- **Webpage:** https://robertomartinmartin.com/
- **Google Scholar:** https://scholar.google.com/citations?user=XOJE8OEAAAAJ
- **Bio:** Roberto Martin-Martin is an Assistant Professor of Computer Science at the University of Texas at Austin. His research integrates robotics, computer vision, and machine learning. He is interested in creating machines that can perceive their environment to acquire task-relevant information, learn and plan a course of action towards a desired new environment configuration, and execute the plan safely and robustly, even under uncertainty and noisy actuation. Previously, He worked as an AI Researcher at Salesforce AI and a Postdoctoral scholar at the Stanford Vision and Learning Lab with Professors Silvio Savarese and Fei-Fei Li.

Ahmad Rushdi (Stanford University)

- **Email:** rushdi@stanford.edu
- **Webpage:** https://aarushdi.github.io
- **Google Scholar:** https://scholar.google.com/citations?user=9F-0uvIAAAAJ
- **Bio:** Ahmad Rushdi is a senior research manager and technical lead at Stanford University's Institute for Human-Centered Artificial Intelligence (HAI). His expertise and research interests are in Uncertainty Quantification (UQ) and Scientific Machine Learning (SciML), with a focus on AI/ML trust and privacy. Prior to HAI, he worked at Sandia National Laboratories and Northrop Grumman, with research experience focusing on machine learning, high-dimensional modeling, and AI applications. Ahmad has a Ph.D. in Electrical and Computer Engineering from UC Davis, and is an active member of IEEE, ACM, and SIAM. Ahmad has extensive experience speaking at and organizing AI-related events, including speaking at a NeurIPS workshop in 2023 and co-organizing a number of workshops, summits, and conferences at Stanford.

Marc Schlichting (Stanford University)

- **Email:** mschl@stanford.edu
- **Webpage:** https://profiles.stanford.edu/marc-schlichting
- **Google Scholar:** https://scholar.google.com/citations?user=IseulBgAAAAJ
- **Bio:** Marc Schlichting is a PhD candidate in Aeronautics and Astronautics at Stanford's Intelligent Systems Laboratory and an M.S. candidate in Neuroscience at Stanford University. His research spans data-driven modeling of glial-neuronal interactions to uncover biologically plausible learning mechanisms, as well as the safety validation of critical systems in areas like precision medicine and robotics. Marc has received several distinctions, including the 2022 Nicholas J. Hoff Award for outstanding master's students and the 2023 Centennial Teaching Assistant Award. He holds an M.S. in Aeronautics and Astronautics from Stanford and a B.S. in Aerospace Engineering from the University of Stuttgart, Germany. Before joining Stanford, he was a visiting researcher at the Institute of Robotics at Georgia Tech.

Peter Stone (The University of Texas at Austin)

- **Email:** pstone@cs.utexas.edu
- **Webpage:** https://www.cs.utexas.edu/~pstone/index.shtml
- **Google Scholar:** https://scholar.google.com/citations?user=qnwjcfAAAAAJ
- **Bio:** Peter Stone is the founder and director of the Learning Agents Research Group (LARG) within the Artificial Intelligence Laboratory in the Department of Computer Science at The University of Texas at Austin, as well as associate department chair and Director of Texas Robotics. He was a co-founder of Cogitai, Inc. and am now Chief Scientist of Sony AI. His main research interest in AI is understanding how we can best create complete intelligent agents. His research focuses mainly on machine learning, multiagent systems, and robotics. His application domains have included robot soccer, autonomous bidding agents, autonomous vehicles, and human-interactive agents.

Hari Subramonyam (Stanford University)

- **Email:** harihars@stanford.edu
- **Webpage:** https://haridecoded.com/
- **Google Scholar:** https://scholar.google.com/citations?user=xG3wqvEAAAAJ
- **Bio:** Hari Subramonyam is a Research Assistant Professor at the Graduate School of Education and Computer Science (by courtesy) at Stanford University. He is also the Ram and Vijay Shriram Faculty Fellow at the Institute for Human-Centered AI (HAI) and a core faculty member of Stanford HCI. His research sits at the intersection of Human-Computer Interaction (HCI) and the Learning Sciences. He has been the lead organizer for workshops at multiple ACM conferences.

Diyi Yang (Stanford University)

- **Email:** diyiy@cs.stanford.edu
- **Webpage:** https://cs.stanford.edu/~diyiy/
- **Google Scholar:** https://scholar.google.com/citations?user=j9jhYqQAAAAJ
- **Bio:** Diyi Yang is an assistant professor in the Computer Science Department at Stanford, affiliated with the Stanford NLP Group, Stanford HCI Group, Stanford AI Lab (SAIL), and Stanford Human-Centered Artificial Intelligence (HAI). She has received a Sloan Research Fellowship (2024), won an NSF CAREER award (2022), and was named Samsung Researcher of the Year (2021). Her research interest is in Socially Aware Natural Language Processing; her research goal is to better understand human communication in social context and build socially aware language technologies to support human-human and human-computer interaction. She has been on the organizing committee for various workshops, including at ACL, NAACL, and EACL, and been an invited speaker for several others, including at ICLR, ICML, and NeurIPS. She has been a Senior Area Chair for ACL, NAACL, and EMNLP and an Area Chair for ICLR, NeurIPS, and other related conferences.

9 Summary and Expected Outcomes

This workshop aims to build the field of HCAI. Generally speaking, we will reframe the narrative around AI systems to emphasize their evolution as collaborators and agents, emphasizing the mutual adaptation of humans and AI to each other through a bidirectional learning process. Through the lens of behavioral evolution and socio-technological norms, we will explore the real-world impacts of AI in socially sensitive domains, investigating how AI systems influence social norms, behaviors, and decision-making processes in ways that challenge traditional views of model development and technology adoption. Furthermore, the workshop will address the evolving nature of bias in learning and decision-making, examining how AI biases emerge, develop, and impact both individual behaviors and societal dynamics. Finally, we will consider the long-term impact on society, exploring how increasing reliance on AI systems will shape social, ethical, and cultural norms in the future.

Participants will leave with a wider view of HAIC concepts, a deeper understanding of ongoing research they would not normally be exposed to, and a professional network in this nascent but rapidly growing research community. Virtual access to workshop materials will be provided. We aim to foster interdisciplinary collaboration by raising awareness of the distinct-yet-related challenges and solutions encountered across various fields and how we can leverage our skills by building such bridges. Our ultimate goal is to produce an overview paper that summarizes the aforementioned concepts, research streams, and open questions in HAIC.

References

A.J. Alvero, Noah Arthurs, Anthony Lising Antonio, Benjamin W. Domingue, Ben Gebre-Medhin, Sonia Giebel, and Mitchell L. Stevens. AI and Holistic Review: Informing Human Reading in College Admissions. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, pp. 200–206, New York NY USA, February 2020. ACM. ISBN 978-1-4503-7110-0. doi: 10.1145/3375627.3375871. URL https://dl.acm.org/doi/10.1145/3375627.3375871.

Jacy Reese Anthis, Kristian Lum, Michael Ekstrand, Avi Feller, Alexander D'Amour, and Chenhao Tan. The Impossibility of Fair LLMs. *Human-Centered Evaluation and Auditing of Language Models*, May 2024.

Ioana Bica, Daniel Jarrett, Alihan Hüyük, and Mihaela van der Schaar. Learning "What-if" Explanations for Sequential Decision-Making. *arXiv:2007.13531 [cs, stat]*, February 2021.

Reuben Binns, Michael Veale, Max Van Kleek, and Nigel Shadbolt. Like trainer, like bot? inheritance of bias in algorithmic content moderation. In Giovanni Luca Ciampaglia, Afra Mashhadi, and Taha

Yasseri (eds.), *Social Informatics*, pp. 405–415, Cham, 2017. Springer International Publishing. ISBN 978-3-319-67256-4.

Su Lin Blodgett, Solon Barocas, Hal Daumé Iii, and Hanna Wallach. Language (Technology) is Power: A Critical Survey of "Bias" in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 5454–5476, Online, 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.485.

Magnus Boman. [No title found]. *Artificial Intelligence and Law*, 7(1):17–35, 1999. ISSN 09248463. doi: 10.1023/A:1008311429414.

Micah Carroll, Davis Foote, Anand Siththaranjan, Stuart Russell, and Anca Dragan. AI Alignment with Changing and Influenceable Reward Functions, May 2024.

Begum Celiktutan, Romain Cadario, and Carey K. Morewedge. People see more of their biases in algorithms. *Proceedings of the National Academy of Sciences*, 121(16):e2317602121, April 2024. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.2317602121.

Alan Chan, Maxime Riché, and Jesse Clifton. Towards the Scalable Evaluation of Cooperativeness in Language Models, March 2023.

Peixin Chang, Shuijing Liu, Tianchen Ji, Neeloy Chakraborty, Kaiwen Hong, and Katherine Rose Driggs-Campbell. A data-efficient visual-audio representation with intuitive fine-tuning for voice-controlled robots. In *Conference on Robot Learning (CoRL)*, 2023.

Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, et al. A survey on evaluation of large language models. *ACM Transactions on Intelligent Systems and Technology*, 15(3):1–45, 2024.

Luisa Damiano and Paul Dumouchel. Anthropomorphism in human–robot co-evolution. *Frontiers in psychology*, 9:468, 2018.

Jeffrey Dastin. Amazon scraps secret ai recruiting tool that showed bias against women. *Reuters*, 2018. URL https://www.reuters.com/article/world/insight-amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK0AG. Accessed: 2024-10-20.

Oliva T. Dias, Dennys M. Antonialli, and Alessandra Gomes. Fighting hate speech, silencing drag queens? artificial intelligence in content moderation and risks to lgbtq voices online. *Sexuality & Culture*, 25(2):700–732, 04 2021. URL https://www.proquest.com/scholarly-journals/fighting-hate-speech-silencing-drag-queens/docview/2495185828/se-2. Copyright - © Springer Science+Business Media, LLC, part of Springer Nature 2020; Last updated - 2024-03-26.

Pierpaolo Donati. Impact of ai/robotics on human relations: co-evolution through hybridisation. *Robotics, AI, and Humanity: Science, Ethics, and Policy*, pp. 213–227, 2021.

Iason Gabriel, Arianna Manzini, Geoff Keeling, Lisa Anne Hendricks, Verena Rieser, Hasan Iqbal, Nenad Tomašev, Ira Ktena, Zachary Kenton, Mikel Rodriguez, Seliem El-Sayed, Sasha Brown, Canfer Akbulut, Andrew Trask, Edward Hughes, A. Stevie Bergman, Renee Shelby, Nahema Marchal, Conor Griffin, Juan Mateos-Garcia, Laura Weidinger, Winnie Street, Benjamin Lange, Alex Ingerman, Alison Lentz, Reed Enger, Andrew Barakat, Victoria Krakovna, John Oliver Siy, Zeb Kurth-Nelson, Amanda McCroskery, Vijay Bolina, Harry Law, Murray Shanahan, Lize Alberts, Borja Balle, Sarah de Haas, Yetunde Ibitoye, Allan Dafoe, Beth Goldberg, Sébastien Krier, Alexander Reese, Sims Witherspoon, Will Hawkins, Maribeth Rauh, Don Wallace, Matija Franklin, Josh A. Goldstein, Joel Lehman, Michael Klenk, Shannon Vallor, Courtney Biles, Meredith Ringel Morris, Helen King, Blaise Agüera y Arcas, William Isaac, and James Manyika. The Ethics of Advanced AI Assistants, April 2024.

Xiao Ge, Chunchen Xu, Daigo Misaki, Hazel Rose Markus, and Jeanne L Tsai. How Culture Shapes What People Want From AI. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, pp. 1–15, Honolulu HI USA, May 2024. ACM. ISBN 9798400703300. doi: 10.1145/3613904.3642660.

Matthias Gerstgrasser, Rylan Schaeffer, Apratim Dey, Rafael Rafailov, Henry Sleight, John Hughes, Tomasz Korbak, Rajashree Agrawal, Dhruv Pai, Andrey Gromov, et al. Is model collapse inevitable? breaking the curse of recursion by accumulating real and synthetic data. *COLM*, 2024.

Thomas Grote and Geoff Keeling. Enabling Fairness in Healthcare Through Machine Learning. *Ethics and Information Technology*, 24(3):39, September 2022. ISSN 1388-1957, 1572-8439. doi: 10.1007/s10676-022-09658-7.

Oliver L. Haimson, Daniel Delmonaco, Peipei Nie, and Andrea Wegner. Disproportionate removals and differing content moderation experiences for conservative, transgender, and black social media users: Marginalization and moderation gray areas. *Proc. ACM Hum.-Comput. Interact.*, 5(CSCW2), October 2021. doi: 10.1145/3479610. URL https://doi.org/10.1145/3479610.

Karen Hao. Ai is sending people to jail – and getting it wrong. *MIT Technology Review*, 2019. URL https://www.technologyreview.com/2019/01/21/137783/algorithms-criminal-justice-ai/. Accessed: 2024-10-20.

Joo-Wha Hong and Dmitri Williams. Racism, responsibility and autonomy in HCI: Testing perceptions of an AI agent. *Computers in Human Behavior*, 100:79–84, November 2019. ISSN 0747-5632. doi: 10.1016/j.chb.2019.06.012.

Abigail Z. Jacobs and Hanna Wallach. Measurement and Fairness. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pp. 375–385, Virtual Event Canada, March 2021. ACM. ISBN 978-1-4503-8309-7. doi: 10.1145/3442188.3445901.

Sanna Järvelä, Guoying Zhao, Janne Heikkilä, Hanna Järvenoja, Kristina Mikkonen, and Satu Kaleva. Hybrid intelligence–human-ai co-evolution and learning in multirealities (hi). In *HHAI 2023: Augmenting Human Intellect*, pp. 392–394. IOS Press, 2023.

Geoff Keeling. Why Trolley Problems Matter for the Ethics of Automated Vehicles. *Science and Engineering Ethics*, 26(1):293–307, February 2020. ISSN 1353-3452, 1471-5546. doi: 10.1007/s11948-019-00096-1.

Geoff Keeling, Katherine Evans, Sarah M. Thornton, Giulio Mecacci, and Filippo Santoni De Sio. Four Perspectives on What Matters for the Ethics of Automated Vehicles. In Gereon Meyer and Sven Beiker (eds.), *Road Vehicle Automation 6*, pp. 49–60. Springer International Publishing, Cham, 2019. ISBN 978-3-030-22932-0 978-3-030-22933-7. doi: 10.1007/978-3-030-22933-7_6.

Zachary Kenton, Ramana Kumar, Sebastian Farquhar, Jonathan Richens, Matt MacDermott, and Tom Everitt. Discovering Agents, August 2022.

Hannah Rose Kirk, Bertie Vidgen, Paul Röttger, and Scott A Hale. Personalisation within bounds: A risk taxonomy and policy framework for the alignment of large language models with personalised feedback. *arXiv preprint arXiv:2303.05453*, 2023.

Jon Kleinberg, Jens Ludwig, Sendhil Mullainathan, and Cass R. Sunstein. Algorithms as discrimination detectors. *Proceedings of the National Academy of Sciences*, pp. 201912790, July 2020. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.1912790117.

Allison Koenecke, Andrew Nam, Emily Lake, Joe Nudell, Minnie Quartey, Zion Mengesha, Connor Toups, John Rickford, Dan Jurafsky, and Sharad Goel. Racial disparities in automated speech recognition. *Proceedings of the National Academy of Sciences*, 117:201915768, 03 2020. doi: 10.1073/pnas.1915768117.

Saurabh Kulkarni and Sunil F Rodd. Context aware recommendation systems: A review of the state of the art techniques. *Computer Science Review*, 37:100255, 2020.

Peter Lee. Learning from Tay's introduction - The Official Microsoft Blog. https://blogs.microsoft.com/blog/2016/03/25/learning-tays-introduction/, 2016.

Ryan Liu, Theodore Sumers, Ishita Dasgupta, and Thomas L. Griffiths. How do Large Language Models Navigate Conflicts between Honesty and Helpfulness? In *Forty-First International Conference on Machine Learning*, June 2024a.

Shuijing Liu, Peixin Chang, Haonan Chen, Neeloy Chakraborty, and Katherine Driggs-Campbell. Learning to navigate intersections with unsupervised driver trait inference. In *2022 International Conference on Robotics and Automation (ICRA)*, pp. 3576–3582, 2022.

Shuijing Liu, Peixin Chang, Zhe Huang, Neeloy Chakraborty, Kaiwen Hong, Weihang Liang, D Livingston McPherson, Junyi Geng, and Katherine Driggs-Campbell. Intention aware robot crowd navigation with attention-based interaction graph. In *IEEE International Conference on Robotics and Automation (ICRA)*, pp. 12015–12021, 2023.

Shuijing Liu, Aamir Hasan, Kaiwen Hong, Runxuan Wang, Peixin Chang, Zachary Mizrachi, Justin Lin, D. Livingston McPherson, Wendy A. Rogers, and Katherine Driggs-Campbell. Dragon: A dialogue-based robot for assistive navigation with visual language grounding. *IEEE Robotics and Automation Letters*, 9(4):3712–3719, 2024b.

Zihan Liu, Han Li, Anfan Chen, Renwen Zhang, and Yi-Chieh Lee. Understanding Public Perceptions of AI Conversational Agents: A Cross-Cultural Analysis. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, pp. 1–17, Honolulu HI USA, May 2024c. ACM. ISBN 9798400703300. doi: 10.1145/3613904.3642840.

Bethanie Maples, Merve Cerit, Aditya Vishwanath, and Roy Pea. Loneliness and suicide mitigation for students using GPT3-enabled chatbots. *npj Mental Health Research*, 3(1):4, January 2024. ISSN 2731-4251. doi: 10.1038/s44184-023-00047-6.

Charles T. Marx, Flavio Du Pin Calmon, and Berk Ustun. Predictive multiplicity in classification. In *Proceedings of the 37th International Conference on Machine Learning*, ICML'20. JMLR.org, 2020.

Yutaka Matsubara, Akihisa Morikawa, Daichi Mizuguchi, and Kiyoshi Fujiwara. Toward human-centered ai framework: An introduction to ai2x co-evolution project. In *SAFECOMP 2023, Position Paper*, 2023.

Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A survey on bias and fairness in machine learning. *ACM computing surveys (CSUR)*, 54(6):1–35, 2021.

Kiana Jafari Meimandi, Matthew Bolton, and Peter Beling. Rl-hat: A new framework for understanding human-agent teaming. In *Proceedings of the AAAI Symposium Series*, volume 1, pp. 80–85, 2023.

Kiana Jafari Meimandi, Matthew L Bolton, and Peter A Beling. Action over words: Predicting human trust in ai partners through gameplay behaviors. *IEEE International Conference on Robot & Human Interactive Communication*, 2024.

Zion Mengesha, Courtney Heldreth, Michal Lahav, Juliana Sublewski, and Elyse Tuennerman. "i don't think these devices are very culturally sensitive."—impact of automated speech recognition errors on african americans. *Frontiers in Artificial Intelligence*, 4, 11 2021. doi: 10.3389/frai.2021.725911.

Hussein Mozannar, Jimin J. Lee, Dennis Wei, Prasanna Sattigeri, Subhro Das, and David Sontag. Effective Human-AI Teams via Learned Natural Language Rules and Onboarding, November 2023.

OpenAI. GPT-4 Technical Report, March 2023.

Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. Generative Agents: Interactive Simulacra of Human Behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, pp. 1–22, San Francisco CA USA, October 2023. ACM. ISBN 9798400701320. doi: 10.1145/3586183.3606763.

Byron Reeves and Clifford Ivar Nass. *The Media Equation: How People Treat Computers, Television, and New Media Like Real People and Places.* CSLI Publications ; Cambridge University Press, Stanford, Calif. : New York, 1996. ISBN 978-1-57586-052-7.

Anka Reuel, Patrick Connolly, Kiana Jafari Meimandi, Shekhar Tewari, Jakub Wiatrak, Dikshita Venkatesh, and Mykel Kochenderfer. Responsible ai in the global context: Maturity model and survey. *arXiv preprint arXiv:2410.09985*, 2024a.

Anka Reuel, Amelia Hardy, Chandler Smith, Max Lamparth, Malcolm Hardy, and Mykel J. Kochenderfer. BetterBench: Assessing AI Benchmarks, Uncovering Issues, and Establishing Best Practices, 2024b. URL https://betterbench.stanford.edu.

Ido Roll and Ruth Wylie. Evolution and Revolution in Artificial Intelligence in Education. *International Journal of Artificial Intelligence in Education*, 26(2):582–599, June 2016. ISSN 1560-4292, 1560-4306. doi: 10.1007/s40593-016-0110-3.

Ilia Shumailov, Zakhar Shumaylov, Yiren Zhao, Nicolas Papernot, Ross Anderson, and Yarin Gal. Ai models collapse when trained on recursively generated data. *Nature*, 631(8022):755–759, 2024.

Taylor Sorensen, Jared Moore, Jillian Fisher, Mitchell Gordon, Niloofar Mireshghallah, Christopher Michael Rytting, Andre Ye, Liwei Jiang, Ximing Lu, Nouha Dziri, Tim Althoff, and Yejin Choi. A Roadmap to Pluralistic Alignment, August 2024.

Hari Subramonyam, Roy Pea, Christopher Pondoc, Maneesh Agrawala, and Colleen Seifert. Bridging the Gulf of Envisioning: Cognitive Challenges in Prompt Based Interactions with LLMs. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, pp. 1–19, Honolulu HI USA, May 2024. ACM. ISBN 9798400703300. doi: 10.1145/3613904.3642754.

Michael Henry Tessler, Michiel A. Bakker, Daniel Jarrett, Hannah Sheahan, Martin J. Chadwick, Raphael Koster, Georgina Evans, Lucy Campbell-Gillingham, Tantum Collins, David C. Parkes, Matthew Botvinick, and Christopher Summerfield. AI can help humans find common ground in democratic deliberation. *Science*, 386(6719):eadq2852, October 2024. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.adq2852.

Aditya Nrusimha Vaidyam, Hannah Wisniewski, John David Halamka, Matcheri S. Kashavan, and John Blake Torous. Chatbots and Conversational Agents in Mental Health: A Review of the Psychiatric Landscape. *The Canadian Journal of Psychiatry*, 64(7):456–464, July 2019. ISSN 0706-7437, 1497-0015. doi: 10.1177/0706743719828977.

Rose Wang and Dorottya Demszky. Is ChatGPT a good teacher coach? measuring zero-shot performance for scoring and providing actionable insights on classroom instruction. In Ekaterina Kochmar, Jill Burstein, Andrea Horbach, Ronja Laarmann-Quante, Nitin Madnani, Anaïs Tack, Victoria Yaneva, Zheng Yuan, and Torsten Zesch (eds.), *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*, pp. 626–667, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.bea-1.53. URL https://aclanthology.org/2023.bea-1.53.

Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Beibin Li, Erkang Zhu, Li Jiang, Xiaoyun Zhang, Shaokun Zhang, Jiale Liu, Ahmed Hassan Awadallah, Ryen W. White, Doug Burger, and Chi Wang. AutoGen: Enabling Next-Gen LLM Applications via Multi-Agent Conversation, October 2023.

Diyi Yang, Tanmay Sinha, David Adamson, and Carolyn Penstein Rosé. "Turn on, tune in, drop out": Anticipating student dropouts in massive open online courses. In *Advances in Neural Information Processing Systems*. NIPS Foundation, 2013.

Diyi Yang, Miaomiao Wen, Iris Howley, Robert Kraut, and Carolyn Rose. Exploring the Effect of Confusion in Discussion Forums of Massive Open Online Courses. In *Proceedings of the Second (2015) ACM Conference on Learning @ Scale*, pp. 121–130, Vancouver BC Canada, March 2015. ACM. ISBN 978-1-4503-3411-2. doi: 10.1145/2724660.2724677.

Alice Qian Zhang, Ryland Shaw, Jacy Reese Anthis, Ashlee Milton, Emily Tseng, Jina Suh, Lama Ahmad, Ram Shankar Siva Kumar, Julian Posada, Benjamin Shestakofsky, Sarah T. Roberts, and Mary L. Gray. The Human Factor in AI Red Teaming: Perspectives from Social and Collaborative Computing, September 2024.

Wenting Zhao, Xiang Ren, Jack Hessel, Claire Cardie, Yejin Choi, and Yuntian Deng. WildChat: 1M ChatGPT Interaction Logs in the Wild, May 2024.