ICLR 2025 Workshop on Bidirectional Human-AI Alignment (Bi-Align @ ICLR 2025 Workshop Proposal)

Hua Shen, Ziqiao Ma, Reshmi Ghosh, Tiffany Knearem
Michael Liu, Tongshuang Wu, Andrés Monroy-Hernández, Diyi Yang, Antoine Bosselut
Furong Huang, Tanu Mitra, Joyce Chai, Marti A. Hearst, Dawn Song, Yang Li

Workshop Summary

The rapid advancements in general-purpose AI has precipitated the urgent need to align these systems with values, ethical principles, and goals of individuals and society at large. This need, commonly referred to as "AI alignment," [26, 25] is crucial for ensuring that AI systems function in a manner that is not only effective but also consistent with human values, minimizing harm and maximizing societal benefits. Traditionally, AI alignment has been viewed as a static, one-way process, with a primary focus on shaping AI systems to achieve desired outcomes and prevent negative side effects [10, 17, 20]. However, as AI systems become more integrated into everyday life and take on more complex decision-making roles, this unidirectional approach is proving inadequate [3]. AI systems interact with humans in evolving, unpredictable ways, generating feedback loops that influence both AI behavior and human responses. This **dynamic interaction necessitates a shift in how we think about alignment**—one that recognizes the bidirectional and adaptive nature of human-AI relationships [22]. Rather than a one-time process or static goal, we should consider alignment as a continuous, evolving engagement between humans and AI, requiring constant reassessment and recalibration.

Moving Toward a Bidirectional Alignment Framework. The concept of *Bidirectional Human-AI Alignment* offers a paradigm shift in how we approach the challenge of human-AI alignment [22], which emphasizes the **dynamic, mutual alignment process**. Particularly, it not only involves an AI-centered perspective, focuses on integrating human specifications into training, steering, and customizing AI. Also, it takes a human-centered perspective into account, aiming to preserve human agency and empower people to think critically when using AI, collaborate effectively with it, and adapt societal approaches to maximize its benefits for humanity. In this way, alignment becomes an interactive, reciprocal process. This ongoing dialogue between humans and machines is essential to achieving true alignment, as it allows both parties to evolve in response to changing contexts, goals, and ethical considerations.

Workshop Goals. To work towards bidirectional human-AI alignment, grounded on our framework from a systematic survey of over 400 alignment papers [22], the core workshop objectives are twofold: (1) broadening the current understanding of AI alignment by focusing on adapting AI to dynamic human needs and societal contexts through research on interactive machine learning, human-in-the-loop learning, and safe reinforcement learning. (2) fostering interdisciplinary collaboration between researchers in multi-disciplinary domains, such as AI, HCI, and social sciences, creating a platform for exchange and innovation. Consequently, this workshop offers a comprehensive perspective and invites researchers to explore diverse aspects of alignment.

A Joint Bi-Align Workshop at CHI 2025 to Bridge HCI+AI Interdisciplinary Gaps. To foster interdisciplinary collaboration on bidirectional human-AI alignment, we propose a complementary Bi-Align workshop at CHI 2025, one of the top Human-Computer Interaction (HCI) conferences, alongside this ICLR 2025 workshop. While the ICLR workshop will focus on AI-centered alignment, the CHI workshop will emphasize human-centered perspectives. To promote cross-disciplinary engagement, participants can attend both workshops—either virtually or in person—and join a shared Slack platform for ongoing interaction. By connecting the HCI and AI/ML communities, we aim to build a collaborative platform that explores bidirectional alignment from diverse academic and practical viewpoints.

Differences from Previous Related Workshops. In the past two years, several workshops have touched upon themes related to our focus, including the Pluralistic Alignment Workshop at NeurIPS 2024, Models of Human Feedback for AI Alignment Workshop at ICML 2024, Representational Alignment Workshop at ICLR 2024. Our workshop stands out from prior workshop and paper efforts in three key ways. First, we broaden the traditional concept of static, uni-directional "AI alignment" by introducing "bidirectional human-AI alignment," which emphasizes the **mutual and dynamic nature** of the alignment process. This expansion is grounded in an extensive review of the relevant literature [22]. Second, our workshop invites submissions that address less-explored alignment challenges with an emphasis on the interdisciplinary perspective, such as how AI influences individuals and society and, further, how to incorporate these dynamics into alignment modeling. Finally, our workshop also has a sister session organized at the CHI 2025 conference focusing on exploring alignment from the human-centered perspective. We particularly allow all our participants to join both workshops in person or virtually, aiming to bridge the communication gaps of both domains and provide them with a shared platform for future connection and collaboration.

Workshop Scope and Topics. This workshop aims to explore the design space from a comprehensive view of bidirectional human-AI alignment, synthesizing and conceptualizing several key topics critical for alignment research:

- **Scope: Broadening the Definition of Alignment.** This topic invites papers of thought-provoking arguments on how to redefine or expand the concept of alignment beyond current boundaries. For instance, how can we incorporate context-aware (situated) alignment or behavioral alignment for LLM agents? How can we better account for evolving human values over time? How to explore the mechanistic alignment of AI and human cognition?
  - *Key Phrases & Example Papers*: value/preference alignment, behavioral alignment, situated alignment, mechanistic alignment, dynamic and lifelong alignment, alignment with human-AI co-evolvement, etc. [3, 1, 23]
- **Opinions: Position Papers and Roadmaps for Future Alignment Research.** This topic invites position papers that explore bold, forward-thinking ideas, theoretical perspectives, or critiques that open new directions for research and practice in this field. These papers need not present finalized or empirically verified results but should stimulate discussion on how we can rethink and advance alignment between humans and AI systems.
  - *Key Phrases & Example Papers*: design principles, roadmap, surveys, literature review, envisioning alignment, challenges and future directions, etc. [22, 11, 24]
- **Specification: Representing Human Values and Norms for AI Alignment.** This topic invites exploratory frameworks or methodologies for formalizing abstract human values and societal and cultural norms into actionable specifications for AI systems, such as the challenges and opportunities in value specification, annotation practices on pluralistic values, especially under disagreements and with geographic diversity, or ways to bridge ethical risks and technical specifications.
  - *Key Phrases & Example Papers*: specifying human objectives, reward hacking and modeling, annotation of human values, etc. [5, 6, 16, 18]
- **Methods: Machine Learning for Aligning AI with Humans.** This topic invites papers to pre/post-train AI models that align with general, broadly shared human values as captured in large-scale datasets or interactive learning processes. We welcome theoretic and practical contributions from relevant machine learning areas like human-in-the-loop learning, multi-task learning, meta-learning, multi-objective reinforcement learning, etc.
  - *Key Phrases & Example Papers*: alignment at scale, post-training, human-in-the-loop learning, multi-task learning, meta-learning, etc. [14, 19, 9]
- **Evaluation: Benchmarks and Metrics for Steerable and Multi-objective AI Alignment.** This topic centers on evaluating and assessing the alignment of AI systems with diverse human values/preferences or behaviors. We invite contributions on benchmarks, as well as innovative evaluation protocols and metrics that address multiple, sometimes conflicting objectives, and explore the steerability of pre-trained models.
  - *Key Phrases & Example Papers*: human-in-the-loop evaluation, steerability, pluralistic value metrics, alignment evaluation framework and protocols etc. [13, 4, 2]
- **Deployment: Customizable Alignment, Interpretability, and Scalable Oversight.** This topic addresses the need to tailor AI systems to specific cultural, societal, or individual values. Also, it explores how to interpret, oversee, and calibrate AI alignment at large-scale deployment. We welcome submissions from relevant areas like continual/life-long learning, interpreting alignment, inference time learning, AI customization, and more.
  - *Key Phrases & Example Papers*: scalable oversight, customization/personalization, interpreting alignment, continual/life-long learning, inference time learning, online learning, situated interaction, etc. [8, 29, 15, 12, 21]
- **Societal Impact and Policy: Fostering An Inclusive Human-AI Alignment Ecosystem.** Human-AI alignment occurs within a broader ecosystem involving multiple stakeholders, including researchers, policymakers, developers, and end-users. This topic explores how to create a collaborative environment where all parties can help shape AI systems that adhere to ethical and technical standards, and the dynamic co-evolvement of AI and human society.
  - *Key Phrases & Example Papers*: cognitive impacts and perspectives on alignment, governance frameworks, feedback loops for continuous alignment, mechanisms for maintaining alignment as AI systems evolve, social impact and AI policy, dynamic impacts of co-evolving alignment, etc. [7, 27, 28, 30]

Tentative Schedule

Tentative Important Dates. We propose to organize this **one-day** workshop schedule with invited talks, poster sessions, a panel session, as well as short contributed talks on oral and outstanding paper submissions.

- Submission Open: 15 October 2024 (Anywhere on Earth)
- Submission Deadline: 1 Feburary 2025 (Anywhere on Earth)
- Notification of Acceptance: 1 March 2025 (Anywhere on Earth)
- Camera Ready Deadline: 1 April 2025 (Anywhere on Earth)
- Workshop Day: 27/28 August 2024 (co-located with ICLR 2025)

Tentative Workshop Schedule. We propose a single-day workshop, from 8:45 AM to 18:00 PM local time (including breaks), in a hybrid format. While we encourage in-person attendance, synchronous online access will be provided. The tentative workshop schedule is detailed in Table 1. We will dedicate sufficient time for group discussions and knowledge sharing; for example, through paper presentations, an expert panel discussion, invited talks, and a paper award ceremony. Our overarching goal is to support participants to meaningfully connect with others in the blooming AI alignment community and to learn from each other. The final workshop schedule will be aligned with the official ICLR 2025 schedule to ensure smooth coordination and integration with the main conference.

Slot	**Theme**
08:50 – 09:00 (15min)	Welcome and Opening Remarks
09:00 – 09:30 (30min)	Keynote 1: Been Kim, Google DeepMind (confirmed)
09:30 – 10:00 (30min)	Keynote 2: Dan Bohus, Microsoft Research (confirmed)
10:00 – 10:30 (30min)	Keynote 3: Frauke Kreuter, LMU Munich and UMD (confirmed)
10:30 – 11:30 (60min)	Poster and Discussion Session (Concurrent Coffee break)
11:30 – 12:30 (60min)	Spotlight Paper Sessions 1 (10min × 6)
12:30 – 13:30 (60min)	Lunch break
13:30 – 14:00 (30min)	Keynote 4: Richard Ngo, OpenAI (confirmed)
14:00 – 14:30 (30min)	Keynote 5: Hung-Yi Lee, National Taiwan University (confirmed)
14:30 – 15:00 (30min)	Keynote 6: Pavel Izmailov, Anthropic/NYU (confirmed)
15:00 – 16:00 (60min)	Poster and Discussion Session (Concurrent Coffee break)
16:00 – 17:00 (60min)	Panel discussion with experts from diverse and well-balanced domains
17:00 – 17:40 (40min)	Spotlight Paper Session 2 (10min × 4)
17:40 – 18:00 (20min)	Paper Award Announcement and Closing Remarks

Table 1: Tentative schedule for the proposed one-day workshop.

Details on the Keynote Sessions. We have commitments from 8 (6 for the proposed ICLR workshop session) prominent researchers and professors with diverse expertise aiming to introduce their perspectives on alignment from diverse fields. We've allotted 25 minutes for each speaker to give their talk, followed by a 5-minute Q&A discussion.

Details on the Panel Session. The discussion panel will include experts with balanced perspectives from academia and industry who will touch on alignment-related workshop topics such as model specification, interpretability, social impacts, human-robotic interaction, etc. When authors are creating their submissions, we will ask them to leave questions to the panel optionally. We will create official accounts on social network platforms such as X (fka Twitter) to publicize the workshop while collecting questions and feedback. During the panel session, we will initiate discussions from the highlighted questions we have collected, and transition to the questions from the live audience.

Details on Paper Presentation Session. Participants will have the opportunity to share their accepted work through either a spotlight paper presentation or a poster. The format for each work will be decided by the organizers and will be based on quality and relevance, with exceptional submissions given preference for a spotlight. We plan to have two spotlight paper sessions, in the morning and afternoon, respectively. Each spotlight session will consist of multiple 8 minute lightning talks, concluded by a 12 minute overall Q&A opportunity. There will be two poster sessions which will run concurrent with a conference coffee break in the morning and afternoon, respectively. All participants will be asked to (optionally) pre-record videos (spotlight papers at 5-7 minute and non-spotlight papers at 1-3 minute lengths). The organizers and program committee will select 2 outstanding paper awards and 1 social impact award and announce them before closing remarks.

Invited Speakers/Panelists

Been Kim (Senior Research Scientist, Google DeepMind) [confirmed]

Bio: Been Kim is a renowned Senior Staff Research Scientist at Google DeepMind, specializing in interpretable machine learning and AI alignment. She pioneered the development of "concept-based" explanations like TCAV, enabling AI systems to provide human-understandable reasoning. Dr. Kim's research has significantly advanced the field of AI interpretability, aiming to bridge the gap between complex AI models and human understanding, making her an influential figure in the field of AI ethics and transparency. She serves on the ICLR board and leadership roles as General Chair, Senior Program Chair, and Workshop Chair at ICLR. Additionally, she has been a steering committee and area chair at FAccT. Dr. Kim gave the keynote at ICLR 2022, ECML 2020, and more.

Research and Talk Topics: Interpretability and Alignment

Frauke Kreuter (Professor and Chair, LMU Munich and University of Maryland) [confirmed]

Bio: Frauke Kreuter is Co-Director of the Social Data Science Center and Professor in the Joint Program in Survey Methodology at the University of Maryland, USA; and Chair of Statistics and Data Science at the Ludwig-Maximilians-University of Munich. She is an elected fellow of the American Statistical Association and the 2020 recipient of the Warren Mitofsky Innovators Award of the American Association for Public Opinion Research. Dr. Kreuter is the Founder of the International Program for Survey and Data Science; Co-founder of the Coleridge Initiative, whose goal is to accelerate data-driven research and policy around human beings and their interactions for program management, policy development, and scholarly purposes.

Research and Talk Topics: Dynamic Human Values, Preferences, and Social Norms

Dan Bohus (Senior Principal Researcher, Microsoft Research) [confirmed]

Bio: Dan Bohus is a Senior Principal Researcher at Microsoft Research, specializing in computational models for multimodal, physically situated interaction. His work focuses on creating systems that reason about their surroundings and seamlessly participate in interactions and collaborations with people in the physical world. Bohus has led groundbreaking research on developing interactive systems that support collaboration and communication between humans and AI, with applications in areas like human-robot interactive systems, embodied conversational agents, intelligent spaces, AR/VR, etc. He gave the keynote at SIGDial, and invited talks at MSR Cambridge AI school, ASRU, and more.

Research and Talk Topics: Multimodal Situated Interaction

Richard Ngo (Research Scientist, OpenAI) [confirmed]

Bio: Richard Ngo is a Research Scientist working on the Governance team at OpenAI, focusing on long-term AI safety and alignment. He was previously a research engineer on the AGI safety team at DeepMind. His work explores the development of general intelligence and the challenges of ensuring that advanced AI systems act in ways aligned with human values. He has written extensively on topics such as AI governance, ethics, and the societal impacts of AI. Richard is a prominent voice in the field, regularly contributing to discussions on AI policy and safety, and has a background in philosophy and computer science. He created the My Alignment Fundamentals Curriculum.

Research and Talk Topics: AI Safety and Model Specification

Pavel Izmailov (Research Scientist / Assistant Professor, Anthropic / New York University) [confirmed]

Bio: Pavel Izmailov is a researcher at Anthropic and an upcoming Assistant Professor at New York University. His research focuses on LLM reasoning, AI for science, and AI alignment. Previously, he worked on reasoning and problem-solving in language models at OpenAI. He contributed to the recent OpenAI o1 models, a new state-of-the-art in LLM reasoning. He have also worked on weak-to-strong-generalization on the superalignment team under Jeff Wu, Jan Leike and Ilya Sutskever. He also had a short stint at xAI reporting to Elon Musk.

Research and Talk Topics: LLM Reasoning for Alignment and AI for Science

Hung-yi Lee (Associate Professor, National Taiwan University) [confirmed]

Bio: Hung-yi Lee is an Associate Professor in the Department of Electrical Engineering at National Taiwan University (NTU), specializing in speech and language processing. His research focuses on deep learning, machine learning, and their applications in speech recognition, natural language processing, and AI education. He has been an influential educator and researcher for his engaging online AI courses, contributing to both the academic community and AI education globally. He gave extensive tutorials, open courses, and invited talks at various conferences and institutes, such as ICASSP, MSR, Google, MIT CSAIL, and more.

Research and Talk Topics: Alignment in Spoken Language Models

Elizebeth Churchill (Department Chair and Professor of HCI, MBZUAI) [confirmed, CHI 2025 session]

Bio: Elizabeth F. Churchill is the department chair and professor of HCI at MBZUAI, previously the Director of User Experience at Google, where she focuses on the intersection of human-computer interaction, UX design, and social computing. With a background in psychology and cognitive science, her research spans areas such as interaction design, AI systems, and human-centered computing. Churchill is a recognized thought leader in HCI, having published extensively and contributed to the design of user experiences that enhance human collaboration with technology.

Research and Talk Topics: User Experience in Alignment

Brad Myers (Director and Professor of HCII, CMU) [confirmed, CHI 2025 session]

Bio: Brad A. Myers is the Charles M. Geschke Director of the Human-Computer Interaction Institute and Professor in the School of Computer Science at Carnegie Mellon University, with an affiliated faculty appointment in the Software and Societal Systems Department. He received the ACM SIGCHI Lifetime Achievement Award in Research, and was awarded the 2022 Alan J. Perlis Award for Imagination in Computer Science. He is an IEEE Life Fellow, ACM Fellow, member of the CHI Academy, and winner of 19 Best Paper type awards and 6 Most Influential Paper Awards. He is the author or editor of over 550 publications, and he has been on the editorial board of 8 journals.

Research and Talk Topics: Interaction Techniques for Alignment

Organizers and Biographies

Hua Shen (Postdoctoral Scholar, University of Washington, `huashen@uw.edu`)

Bio: Hua Shen is a postdoctoral scholar at the University of Washington. Her research centers on bidirectional human-AI alignment, covering HCI and various AI fields, including NLP, speech processing, and computer vision. She received multiple awards, including AIED'24 Best Paper, CSCW'23 Best Demo, IUI'23 Best Paper Honorable Mention, 2023 Google Research Science Conference Scholarships, and 2023 Rising Stars of Data Science. She also served as Associate Chairs for CHI, CHI LBW, Program Committees for ACL, EMNLP, and more. She earned her Ph.D. from Penn State University and completed a postdoctoral fellowship at the University of Michigan. She also served as MichiganAI Seminar Tsar, and Co-organizer of Michigan Interactive and Social Computing group.

Ziqiao Ma (PhD Candidate, University of Michigan, `marstin@umich.edu`)

Bio: Martin Ziqiao Ma is a Ph.D. candidate at the University of Michigan. His research stands on the intersection of language, interaction, and embodiment from a cognitive perspective, with the goal of grounding and aligning language agents to non-linguistic modalities and rich interactive contexts. He is a recipient of the Weinberg Cognitive Science Fellowship, an Outstanding Paper Award at ACL 2023, and the Amazon Alexa Prize Simbot Challenge Award. He co-organized the 4th SpLU-RoboNLP workshop at ACL 2024, served as the poster/demo chair for the 5th Michigan AI Symposium 2022, and as a regular program committee for various ML/NLP venues.

Reshmi Ghosh (Applied Scientist Lead, Microsoft, `reshmighosh@microsoft.com`)

Bio: Reshmi Ghosh is an Applied Scientist Lead for GenAI Safety in Microsoft's Responsible AI and Security team and has recently released novel methods for LLM Safety for 1P and 3P use. She was also the core architect in designing M365 CoPilots in 2023, and has previously worked on integrating machine learning features to Excel, Word, and PowerPoint. She graduated with a Ph.D. in data reconstruction using NLP methods for mitigating climate change, from Carnegie Mellon University in 2021, and is a research advisor for teams in MIT CSAIL, UMass Amherst, UCLA, and Oxford University. She has published in EMNLP, ICML, NeurIPS, ACL, ACM CIKM, KDD, etc.

Tiffany Knearem (User Experience Researcher, Google, `tknearem@google.com`)

Bio: Tiffany Knearem is a User Experience Researcher on the Material Design team at Google. Her research focus is on product designer-developer collaboration, creativity support tooling and opportunities for AI in the user interface (UI) design space. She holds a PhD in Information Sciences and Technologies with emphasis on Human-Computer Interaction from Pennsylvania State University, advised by Dr. John M. Carroll. She co-organized the CHI 2024 workshop on Computational UI.

Michael Xieyang Liu (Research Scientist, Google DeepMind, `lxieyang@google.com`)

Bio: Michael Xieyang Liu is a research scientist at Google DeepMind. His research aims to improve human-AI interaction, with a particular focus on human interaction with multimodal large language models and controllable AI. Michael organized the Sensemaking workshop at CHI 2024. Michael previously earned his Ph.D. from the Human-Computer Interaction Institute at Carnegie Mellon University. There, he worked at the intersection of HCI, programming tools, sensemaking, intelligent user interfaces, and human-AI interaction, where he designed and built systems that accelerate online sensemaking for developers and facilitate human-AI interactions for end-users.

Tongshuang Wu (Assistant Professor, Carnegie Mellon University, sherryw@cs.cmu.edu)

Bio: Sherry Wu is an Assistant Professor at the Human-Computer Interaction Institute, Carnegie Mellon University. Her research lies at the intersection of Human-Computer Interaction and Natural Language Processing, aiming to design, evaluate, build, and interact with AI systems that are compatible with actual human goals. Sherry has organized three workshops at NLP and HCI conferences: Shared Stories and Lessons Learned workshop at EMNLP 2022 and Trust and Reliance in AI-Human Teams at CHI 2023-2024. She has also given two tutorials related to Human-AI Interaction at EMNLP 2023 and NAACL 2024. Before joining CMU, Sherry received her Ph.D. degree from the University of Washington.

Andrés Monroy-Hernández (Assistant Professor, Princeton University, andresmh@princeton.edu)

Bio: Andrés Monroy-Hernández is an Assistant Professor co-leading the Princeton HCI Lab at Princeton University, where his research focuses on human-computer interaction and social computing. He is also an associated faculty at Princeton's Center for Information Technology and Policy, the Keller Center for Innovation, the DeCenter, the Program in Cognitive Science, and the Program in Latin American Studies. Before Princeton, he founded the HCI research team at Snap and led the FUSE Labs at MSR. He received his Ph.D. degree in Media Arts and Sciences from MIT, was named one of the 35 Innovators under 35 by the MIT Technology Review. He was the technical program co-chair, editor, and steering committee for ACM CSCW conferences.

Diyi Yang (Assistant Professor, Stanford University, diyiy@cs.stanford.edu)

Bio: Diyi Yang is an Assistant Professor in the Computer Science Department at Stanford University, affiliated with the Stanford NLP Group, Stanford HCI Group, Stanford AI Lab (SAIL), and Stanford Human-Centered Artificial Intelligence (HAI). Her research focuses on human-centered natural language processing and computational social science. Diyi has organized four workshops at NLP conferences: Widening NLP Workshops at NAACL 2018 and ACL 2019, Casual Inference workshop at EMNLP 2021, NLG Evaluation workshop at EMNLP 2021, and Shared Stories and Lessons Learned workshop at EMNLP 2022. She also gave a tutorial at ACL 2022 on Learning with Limited Data, and a tutorial at EACL 2023 on Summarizing Conversations at Scale. Diyi and Sherry have co-developed a new course on Human-Centered NLP that has been offered at both Stanford and CMU.

Antoine Bosselut (Assistant Professor, EPFL, antoine.bosselut@epfl.ch)

Bio: Antoine Bosselut is an Assistant Professor in the School of Computer and Communication Sciences at EPFL, specializing in natural language processing (NLP) and machine learning. His research focuses on building knowledge-enhanced language models that can reason and make inferences about the world. Antoine's work aims to bridge the gap between language models and human-like understanding, contributing significantly to areas such as commonsense reasoning and narrative understanding, with applications in improving the reasoning capabilities of AI systems. Previously, he was a postdoctoral researcher at Stanford University working with Jure Leskovec and Chris Manning. He completed a PhD at the University of Washington, working with Yejin Choi. He was named to the Forbes 30 under 30 list in Science & Healthcare.

Furong Huang (Associate Professor, University of Maryland, furongh@umd.edu)

Bio: Furong Huang is an Associate Professor in the Department of Computer Science at the University of Maryland. Specializing in trustworthy machine learning, AI for sequential decision-making, and high-dimensional statistics, Dr. Huang focuses on applying theoretical principles to solve practical challenges in contemporary computing. Her research centers on creating reliable and interpretable machine learning models that operate effectively in real-world settings. She has also made significant strides in sequential decision-making, aiming to develop algorithms that optimize performance and adhere to ethical and safety standards. She organized the NeurIPS competition of "A Stress-Test Challenge for Image Watermarks," chair and organizer of NSF-Amazon Fairness in AI Principle Investigator Meeting, Co-organizer of the NSF-IEEE workshop, and more.

Tanu Mitra (Associate Professor, University of Washington, tmitra@uw.edu)

Bio: Tanu Mitra is an Associate Professor at the Information School at the University of Washington, and co-founded the Responsibility in AI Systems and Experiences (RAISE) Center. Her research blends human-centered data science and social science principles to develop new knowledge, methods, and systems to defend against the epistemic risks of online mis(dis)information, bias, hate and harms. Tanu's work has been supported by grants from the NSF, NIH, DoD, Social Science One, and other Foundations. Her research has been recognized through multiple awards and honors, including an NSF-CRII, an early career ONR-YIP, Adamic-Glance Distinguished Young Researcher award and Virginia Tech College of Engineering Outstanding New Assistant Professor award, along with several best paper awards. Dr. Mitra currently serves on Spotify's safety advisory board and has previously served on the advisory board of the Social Science Research Council's Social Data Initiative.

Joyce Chai (Professor, University of Michigan, chaijy@umich.edu)

Bio: Joyce Chai is a Professor in the Department of Electrical Engineering and Computer Science at the University of

Michigan. Her research interests span NLP and embodied AI to human-AI collaboration. Her current work explores the intersection between language, perception, and action to enable situated communication with embodied agents. She served on the executive board of NAACL and as Program Co-Chair for multiple conferences, most recently ACL 2020. She is a recipient of the NSF Career Award and multiple paper awards with her students (e.g., Best Long Paper Award at ACL 2010, Outstanding Paper Awards at EMNLP 2021 and ACL 2023). She is a Fellow of ACL.

Marti A. Hearst (Professor, University of California, Berkeley, hearst@berkeley.edu)

Bio: Marti A. Hearst is a professor and previously the Interim Dean for the UC Berkeley School of Information. She is both an ACL Fellow and a SIGCHI Academy member, and former ACL President. Her research has long combined HCI and NLP; recent projects include adding interactivity to scholarly documents and creating interactive newspods. She recently gave invited keynote talks at the EACL NLP + HCI workshop, the KDD Workshop on Data Science with a Human in the Loop, and she advised the 2022 NAACL program chairs on the Human-Centered NLP special theme. She has taught courses in NLP, HCI, and information visualization for 25 years.

Dawn Song (Professor, University of California, Berkeley, dawnsong@cs.berkeley.edu)

Bio: Dawn Song is a Professor in the Department of Electrical Engineering and Computer Science at UC Berkeley. Her research interest lies in AI and deep learning, blockchain/web3, security and privacy. She is the recipient of various awards including the MacArthur Fellowship, the Guggenheim Fellowship, the NSF CAREER Award, the Alfred P. Sloan Research Fellowship, the MIT Technology Review TR-35 Award, and several Test-of-Time and Best Paper Awards from top conferences in Computer Security and Deep Learning. She is an ACM Fellow and an IEEE Fellow. She is ranked the most cited scholar in computer security (AMiner Award). She obtained her Ph.D. degree from UC Berkeley. Prior to joining UC Berkeley as a faculty, she was a faculty at Carnegie Mellon University from 2002 to 2007. She is also a serial entrepreneur and has been named on the Female Founder 100 List by Inc. and Wired25 List of Innovators.

Yang Li (Senior Staff Research Scientist, Google DeepMind, liyang@google.com)

Bio: Yang Li is a Senior Staff Research Scientist at Google DeepMind, and an affiliate faculty member at University of Washington. His research lies at the intersection of HCI and AI, focusing on general deep learning research and models for solving human interactive intelligence problems and improving user experiences. He earned a Ph.D. degree in Computer Science from the Chinese Academy of Sciences, and conducted postdoctoral research at UC Berkeley EECS. Yang has extensively published in top venues across both the HCI and ML fields, including CHI, UIST, ICML, ACL, CVPR, NeurIPS, ICLR and KDD, and has constantly served as area chairs or senior area chairs across the HCI and ML fields. Yang is an editor of the Springer book on AI for HCI: A Modern Approach, and an organizer of multiple workshops that bridges the HCI and AI/ML field, including the first AI&HCI workshp at ICML.

Diversity and Inclusion Commitment

Diversity of Organizing Committee. The organizing committee consists of professors and practitioners from a wide variety of demographic backgrounds and experiences, and we aim to promote diversity along several axes, including gender, seniority, experience, geographic locations, affiliation, and research areas. Among the 12 organizers, 7 of them identified themselves as female and 5 as male. There are 9 from academia, including two full professors, 1 associate professor, 4 assistant professors, 1 postdoctoral scholar and 1 Ph.D. candidate. Particularly, we have 3 industrial practitioners from frontier companies, including 1 Applied Scientist Lead for GenAI Safety at Microsoft, 1 User Experience Resaercher and 1 Senior Staff Research Scientist at Google DeepMind. Organizers are affiliated with different institutions covering North America and Europe, including UW, UMich, Microsoft, Stanford, EPFL, UMD, Google DeepMind, CMU, Princeton, UC Berkeley.

Diversity of Speakers. Our commitment of fostering diversity is evident in our choice of invited speakers. They come from varied institutions in both academic and industry, including Google DeepMind, OpenAI, NTU, Microsoft, and Anthropic/NYU. Notably, we invite speakers to introduce their perspectives of alignment from diverse research fields. We're pleased that our current list features speakers who have prominently contributed to diverse fields, including interpretability, model development, spoken LLM, multimodal situated interaction, reasoning, social impact, and beyond. Moreover, we are conscious of gender and geographic location representation, with both female and male speakers from North American and Asian institutions.

Diversity of Participants. Our workshop is committed to fostering an inclusive environment that embraces participants from diverse backgrounds. With a fully representative organizing committee and a lineup of invited speakers, we aim to create a space where attendees can connect over shared perspectives and engage in meaningful discussions.

Workshop Modality, Logistics, and Virtual Access

Workshop Website. We are creating a website to advertise and disseminate the workshop's information. We will also use this website to share workshop contributions, including accepted papers, and support future engagement.

Hybrid Formats. We will host the workshop in a **hybrid format**, primarily in-person with an option for remote participation. All sessions will be live-streamed, following the ICLR 2025 guidelines. We will leverage the conference center's standard equipment to meet the technical needs. A workshop website and Slack platform will serve as central hubs for engagement, offering details such as the call for papers, program schedule, organizers, speakers, and preprints of accepted position papers. Our team has experience with platforms like Zoom and GatherTown Live for organizing such events.

Virtual access to workshop and Asynchronous Materials. We provide asynchronous materials for all participants to access offline through both the workshop website and Slack platform. In case any technical or accessibility issues arise, we provide all important information, such as the program schedule, list of organizers and speakers, and preprints of accepted papers, on the workshop website. Besides, we allow all participants to engage in the workshop Slack for Q&A and discussion. Furthermore, we will release the videos of the workshop presentations on YouTube and list them on our workshop website.

Plans to Publish Workshop Proceedings. Although the accepted papers are non-archival, we plan to compile a comprehensive report detailing the outcome of the event. The workshop papers will be hosted on our workshop website before and available after the workshop, allowing a broad audience to be informed about the content. In addition, we will look for opportunities to curate an edited volume or a special journal issue, where participants will be invited to contribute their work.

Accessibility. To ensure broader participation, the workshop will be held in a hybrid format, as we believe in-person interactions are invaluable for sparking research ideas. However, the entire workshop will also be live-streamed to reach a wider audience. We will prioritize clear and open communication before, during, and after the event to share key insights and expand the conversation. In addition, we will publish the titles of invited talks, speaker details, and PDFs of contributed works on the workshop website ahead of time. Important discoveries and outcomes will be shared widely through regular website updates, video postings, and social media platforms.

Workshop Dissemination, Submission, and Anticipated Size

Workshop Dissemination. Before the workshop, we will distribute a call for participation across a variety of ML-related emailing lists and social media, like Twitter and LinkedIn. The call will invite researchers and practitioners to contribute by submitting research papers. We will also advertise the workshop at upcoming ML-related conferences, among research groups, and through our professional networks. We will leverage the workshop website to help potential participants get familiar with the workshop's scope, goals, and other information about the workshop.

Workshop Submission and Expected Size. We invite both (1) short research papers, up to 4 pages in length, and (2) full-length research papers, up to 9 pages (excluding references and supplementary materials). All papers must be in PDF format and submitted through the OpenReview submission portal, following the ICLR 2025 template. All accepted papers will be designated for poster presentations. In addition, we will select 6-12 papers for short oral presentations. Some papers will receive special recognition, with 2 outstanding paper awards and 1 social impact award. All accepted papers will be available on the workshop homepage and through OpenReview, but will be considered non-archival. The submission to CHI and ICLR workshops will be through separate protocols, and each manuscript can only be committed to one of the venues if accepted to both. Given the scope and multidisciplinary nature of this workshop, we expect an audience size of 300-500.

Program Committee Members

Based on the topic, broad scope, and audience of previous ICLR workshops, we anticipate receiving 50-100 paper submissions. We will ensure that each submission undergoes 3-4 reviews, including thorough ethical reviews, with a maximum of 4 papers assigned per reviewer. Our program committee (PC) will consist of a diverse group of experts in areas such as NLP, RL, HCI, ML, AI/ML Ethics, and more, selected based on their publication record and expertise.

We also plan to invite authors of submitted papers to serve as additional reviewers. Once the submission site is set up, we will confirm the availability of the PC members to participate in the review process. In addition, we will proactively reach out to emergency reviewers to ensure the review process stays on schedule and that all submissions receive timely and thorough evaluations.

We will manage the conflicts of interest in assessing submitted contributions. Particularly, the organizer and program committee member would not be involved in the assessment of a submission from someone with the same organization.

- Nitesh Goyal, Google DeepMind
- Tuhin Chakrabarty, Salesforce
- Jiayi Pan, UC, Berkeley
- Beatriz Borges, EPFL
- Qianou Ma, CMU
- Yongyuan Liang, UMD
- Emily Yuwei Bao, UMich
- Sumit Asthana, UMich
- Ryan Liu, Princeton
- Amna Liaqat, Princeton
- Tal August, UIUC
- Sarah Sterman, UIUC
- Haoyi Qiu, UCLA
- Olivia Simin Fan, EPFL
- Ruoxi Ning, UWaterloo
- Chen Zeming, EPFL
- Gao Silin, EPFL
- Badr Alkhamissy, EPFL

References

[1] Prithviraj Ammanabrolu, Liwei Jiang, Maarten Sap, Hannaneh Hajishirzi, and Yejin Choi. Aligning to social norms and values in interactive narratives. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 5994–6017, 2022.

[2] Noam Benkler, Drisana Mosaphir, Scott Friedman, Andrew Smart, and Sonja Schmer-Galunder. Assessing llms for moral value pluralism. *Workshop on AI meets Moral Philosophy and Moral Psychology: An Interdisciplinary Dialogue about Computational Ethics, NeurIPS 2023*, 2023.

[3] Micah Carroll, Davis Foote, Anand Siththaranjan, Stuart Russell, and Anca Dragan. Ai alignment with changing and influenceable reward functions. *arXiv:2405.17713*, 2024.

[4] Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Banghua Zhu, Hao Zhang, Michael Jordan, Joseph E. Gonzalez, and Ion Stoica. Chatbot arena: An open platform for evaluating LLMs by human preference. In *Forty-first International Conference on Machine Learning*, 2024.

[5] John Joon Young Chung, Wooseok Kim, Kang Min Yoo, Hwaran Lee, Eytan Adar, and Minsuk Chang. Talebrush: Sketching stories with generative pretrained language models. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, pp. 1–19, 2022.

[6] Zibin Dong, Yifu Yuan, Jianye HAO, Fei Ni, Yao Mu, YAN ZHENG, Yujing Hu, Tangjie Lv, Changjie Fan, and Zhipeng Hu. Aligndiff: Aligning diverse human preferences via behavior-customisable diffusion model. In *The Twelfth International Conference on Learning Representations*, 2024.

[7] Yann Dubois, Chen Xuechen Li, Rohan Taori, Tianyi Zhang, Ishaan Gulrajani, Jimmy Ba, Carlos Guestrin, Percy S Liang, and Tatsunori B Hashimoto. Alpacafarm: A simulation framework for methods that learn from human feedback. volume 36, 2023.

[8] Virginia Felkner, Ho-Chun Herbert Chang, Eugene Jang, and Jonathan May. Winoqueer: A community-in-the-loop benchmark for anti-lgbtq+ bias in large language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 9126–9140, 2023.

[9] Zhibin Gou, Zhihong Shao, Yeyun Gong, yelong shen, Yujiu Yang, Nan Duan, and Weizhu Chen. CRITIC: Large language models can self-correct with tool-interactive critiquing. In *The Twelfth International Conference on Learning Representations*, 2024.

[10] Nitesh Goyal, Minsuk Chang, and Michael Terry. Designing for human-agent alignment: Understanding what humans want from their agents. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, pp. 1–6, 2024.

[11] Sayash Kapoor, Rishi Bommasani, Kevin Klyman, Shayne Longpre, Ashwin Ramaswami, Peter Cihon, Aspen Hopkins, Kevin Bankston, Stella Biderman, Miranda Bogen, et al. Position: On the societal impact of open foundation models. In *International Conference on Machine Learning*, pp. 23082–23104. PMLR, 2024.

[12] Preethi Lahoti, Nicholas Blumm, Xiao Ma, Raghavendra Kotikalapudi, Sahitya Potluri, Qijun Tan, Hansa Srinivasan, Ben Packer, Ahmad Beirami, Alex Beutel, et al. Improving diversity of demographic representation in large language models via collective-critiques and self-voting. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 10383–10405, 2023.

[13] Junyi Li, Charith Peris, Ninareh Mehrabi, Palash Goyal, Kai-Wei Chang, Aram Galstyan, Richard Zemel, and Rahul Gupta. The steerability of large language models toward data-driven personas. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 7283–7298, 2024.

[14] Minzhi Li, Taiwei Shi, Caleb Ziems, Min-Yen Kan, Nancy Chen, Zhengyuan Liu, and Diyi Yang. Coannotating: Uncertainty-guided work allocation between human and large language models for data annotation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 1487–1505, 2023.

[15] Jessica Maghakian, Paul Mineiro, Kishan Panaganti, Mark Rucker, Akanksha Saran, and Cheng Tan. Personalized reward learning with interaction-grounded learning (igl). In *The Eleventh International Conference on Learning Representations*, 2023.

[16] Jakob Mökander, Jonas Schuett, Hannah Rose Kirk, and Luciano Floridi. Auditing large language models: a three-layered approach. *AI and Ethics*, pp. 1–31, 2023.

[17] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022.

[18] Savvas Petridis, Benjamin D Wedin, James Wexler, Mahima Pushkarna, Aaron Donsbach, Nitesh Goyal, Carrie J Cai, and Michael Terry. Constitutionmaker: Interactively critiquing large language models by converting feedback into principles. In *Proceedings of the 29th International Conference on Intelligent User Interfaces*, pp. 853–868, 2024.

[19] Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. volume 36, 2023.

[20] Shibani Santurkar, Esin Durmus, Faisal Ladhak, Cinoo Lee, Percy Liang, and Tatsunori Hashimoto. Whose opinions do language models reflect? In *International Conference on Machine Learning*, pp. 29971–30004. PMLR, 2023.

[21] Hua Shen, Yuguang Yang, Guoli Sun, Ryan Langman, Eunjung Han, Jasha Droppo, and Andreas Stolcke. Improving fairness in speaker verification via group-adapted fusion network. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 7077–7081. IEEE, 2022.

[22] Hua Shen, Tiffany Knearem, Reshmi Ghosh, Kenan Alkiek, Kundan Krishna, Yachuan Liu, Ziqiao Ma, Savvas Petridis, Yi-Hao Peng, Li Qiwei, Sushrita Rakshit, Chenglei Si, Yutong Xie, Jeffrey P. Bigham, Frank Bentley, Joyce Chai, Zachary Lipton, Qiaozhu Mei, Rada Mihalcea, Michael Terry, Diyi Yang, Meredith Ringel Morris, Paul Resnick, and David Jurgens. Towards bidirectional human-ai alignment: A systematic review for clarifications, framework, and future directions. *arXiv preprint arXiv:2406.09264*, 2024.

[23] Hua Shen, Tiffany Knearem, Reshmi Ghosh, Yu-Ju Yang, Tanushree Mitra, and Yun Huang. Valuecompass: A framework of fundamental values for human-ai alignment. *arXiv preprint arXiv:2409.09586*, 2024.

[24] Taylor Sorensen, Jared Moore, Jillian Fisher, Mitchell Gordon, Niloofar Mireshghallah, Christopher Michael Rytting, Andre Ye, Liwei Jiang, Ximing Lu, Nouha Dziri, et al. A roadmap to pluralistic alignment. *arXiv:2402.05070*, 2024.

[25] Michael Terry, Chinmay Kulkarni, Martin Wattenberg, Lucas Dixon, and Meredith Ringel Morris. Ai alignment in the design of interactive ai: Specification alignment, process alignment, and evaluation support. *arXiv:2311.00710*, 2023.

[26] Wikipedia. AI alignment — Wikipedia, the free encyclopedia. `http://en.wikipedia.org/w/index.php?title=AI%20alignment&oldid=1220304776`, 2024. [Online; accessed 05-May-2024].

[27] Binfeng Xu, Xukun Liu, Hua Shen, Zeyu Han, Yuhan Li, Murong Yue, Zhiyuan Peng, Yuchen Liu, Ziyu Yao, and Dongkuan Xu. Gentopia. ai: A collaborative platform for tool-augmented llms. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 237–245, 2023.

[28] Yifu Yuan, Jianye Hao, Yi Ma, Zibin Dong, Hebin Liang, Jinyi Liu, Zhixin Feng, Kai Zhao, and Yan Zheng. Uni-rlhf: Universal platform and benchmark suite for reinforcement learning with diverse human feedback. In *The Twelfth International Conference on Learning Representations*, 2024.

[29] Siyan Zhao, John Dang, and Aditya Grover. Group preference optimization: Few-shot alignment of large language models. In *The Twelfth International Conference on Learning Representations*, 2023.

[30] Xuhui Zhou, Hao Zhu, Leena Mathur, Ruohong Zhang, Haofei Yu, Zhengyang Qi, Louis-Philippe Morency, Yonatan Bisk, Daniel Fried, Graham Neubig, et al. Sotopia: Interactive evaluation for social intelligence in language agents. In *In the International Conference on Learning Representations*, 2023.