ICLR 2025 Workshop on Bidirectional Human-AI Alignment

April 27 or 28, 2025 @ Singapore

ICLR

Join our Slack

To foster interdisciplinary research on "**bidirectional human-AI alignment**" (see **definition**), we invite authors from all disciplines (e.g., **ML**, **HCI**, **NLP**, **Vision**, **Speech**, **Social Computing**). This workshop calls for papers with **various alignment topics**, including but not limited to:

- **Scope:** Broadening the Definition of Alignment;
- **Opinion:** Position Papers and Roadmaps for Future Alignment Research;
- **Specification:** Representing Human Values, Behavior, Cognition, Societal Norms for Alignment;
- **Methods:** Algorithms, Interaction Mechanisms, UX Design, RLHF, for Aligning AI with Humans;
- **Evaluation:** Benchmarks, Metrics or Human Evaluation for Multi-objective AI Alignment;
- **Deployment:** Customizable Alignment, Steerability, Interpretability, and Scalable Oversight;
- **Societal Impact and Policy:** Fostering An Inclusive Human-AI Alignment Ecosystem.

We call for **2-page** (tiny), **4-page** (short), and **9-page** (long) papers.

Keynotes & Topics

Been Kim
Google DeepMind
Interpretability & Alignment

Brad Myers
CMU
Interaction for Alignment

Frauke Kreuter
LMU Munich / Univ. of Maryland
Dynamic Human Values & Social Norms

Hung-yi Lee
National Taiwan University
Alignment in Spoken LLMs

Richard Ngo
Prev. OpenAI
AI Safety & Specification

Dan Bohus
Microsoft Research
Multimodal Situated Interaction

Pavel Izmailov
Anthropic / NYU
Reasoning for Alignment

Tentative Important Dates

- **Submission Open:** *January 1, 2025.*
- **Submission Deadline:** *February 3, 2025.*
- **Notification of Acceptance:** *March 3, 2025.*
- **Workshop Day:** *April 27 or 28, 2025.*

All deadlines are AoE.

Organizing Committee

Hua Shen
University of Washington

Ziqiao Ma
University of Michigan

Reshmi Ghosh
Microsoft

Tiffany Knearem
Google

Michael Liu
Google DeepMind

Sherry Wu
CMU

Andrés Monroy-Hernández
Princeton

Diyi Yang
Stanford

Antoine Bosselut
EPFL

Furong Huang
University of Maryland

Tanu Mitra
University of Washington

Joyce Chai
University of Michigan

Marti A. Hearst
UC Berkeley

Dawn Song
UC Berkeley

Yang Li
Google DeepMind