Bidirectional Human-AI Alignment (Bi-Align): Exploring Design, Interaction, Evaluation, and Critical Thinking of Alignment

HUA SHEN, University of Washington, USA
TIFFANY KNEAREM, Google, USA
RESHMI GHOSH, Microsoft, USA
MICHAEL XIEYANG LIU, Google DeepMind, USA
ANDRÉS MONROY-HERNÁNDEZ, Princeton University, USA
TONGSHUANG WU, Carnegie Mellon University, USA
DIYI YANG, Stanford University, USA
TANU MITRA, University of Washington, USA
YANG LI, Google DeepMind, USA
MARTI A. HEARST, University of California, Berkeley, USA

Recent advancements in general-purpose AI have highlighted the critical need to align AI systems with the goals, ethical principles, and values of individuals and groups – a concept widely known as *AI alignment*. Traditionally, this has been approached as a static, one-way process focused on ensuring that AI systems achieve desired outcomes while minimizing negative side effects. However, this unidirectional perspective fails to account for the dynamic and evolving interaction between humans and AI, necessitating a shift toward a bidirectional, interconnected model of human-AI alignment. Based on a systematic survey of over 400 papers, which introduced a conceptual framework for "Bidirectional Human-AI Alignment," [6] this workshop seeks to broaden the design space of human-AI alignment by focusing on aligning humans with AI. This direction aims to support human agency, empower critical thinking in AI use, foster effective human-AI collaboration, and promote societal adaptations to maximize the benefits of AI for humanity. The workshop is part of a joint proposal for "Bidirectional Human-AI Alignment" workshops at both the CHI 2025 and ICLR 2025 conferences. The workshop aims to build a shared platform and bring together experts from HCI, AI, machine learning, psychology, and social sciences to advance interdisciplinary research and collaboration on human-AI alignment.

Additional Key Words and Phrases: bidirectional human-AI alignment, fundamental values, human-AI interaction

ACM Reference Format:
Hua Shen, Tiffany Knearem, Reshmi Ghosh, Michael Xieyang Liu, Andrés Monroy-Hernández, Tongshuang Wu, Diyi Yang, Tanu Mitra, Yang Li, and Marti A. Hearst. 2018. Bidirectional Human-AI Alignment (Bi-Align): Exploring Design, Interaction, Evaluation, and Critical Thinking of Alignment. In . ACM, New York, NY, USA, 9 pages. https://doi.org/XXXXXXX.XXXXXXX

1 MOTIVATION

The rapid advancements in general-purpose AI have brought to the forefront the urgent need to align these systems with the ethical principles, values, and goals of individuals and society at large. This need, commonly referred to as "AI alignment," [8] is crucial for ensuring that AI systems function in a manner that is not only effective but also consistent

Manuscript submitted to ACM

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 [2, 4, 5]. This conventional approach typically frames alignment as a technical challenge—ensuring AI behaves according to human intentions by engineering systems that meet predetermined criteria [10]. However, as AI systems become more integrated into everyday life and take on more complex decision-making roles, this unidirectional approach is proving inadequate [1]. 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 [6]. Rather than treating alignment as a one-time process or static goal, we should consider it 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 [6], which emphasizes the more **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 human and machine is essential to achieving true alignment, as it allows both parties to evolve in response to changing contexts, goals, and ethical considerations.

Expanding AI Alignment with a Human-Centered Perspective. AI alignment has traditionally been viewed as a subset of AI safety [10], often neglecting the significant contributions from HCI and social science fields. This workshop seeks to broaden the design space of human-centered AI alignment, grounded on the Bidirectional Human-AI Alignment framework, by encouraging HCI researchers to explore a range of topics such as alignment design, interaction, evaluation, and critical thinking. While HCI research has contributed to areas like human-AI interaction and explanation, this workshop offers a more comprehensive perspective. It invites researchers to explore diverse aspects of alignment, including fundamental human values, understanding and perception of AI, critical thinking about AI, human-AI collaboration, user experience design, the social impact of AI, and the dynamic evolution of human-AI alignment. Further details are provided in the "Workshop Scope" and "Call for Papers" sections.

A Joint Bi-Align Workshop at ICLR 2025 to Bridge HCI+AI Interdisciplinary Gaps. As bidirectional human-AI alignment necessitates an interdisciplinary effort in achieving true alignment, we propose a joint Bi-Align workshop at ICLR 2025 (one of the top Machine Learning conferences), complementing this CHI 2025 workshop. To distinguish the two workshops as well as accommodate the corresponding participants' expertise and domains, we design the Bi-Align workshop at CHI 2025 (i.e., this proposal) to focus on a human-centered alignment perspective. In comparison, we design the Bi-Align workshop at ICLR 2025 to emphasize a technical focus on AI-centered alignment perspectives. Please see the ICLR Bi-Align workshop proposal at this link. Furthermore, to encourage interdisciplinary communication and collaboration, we allow participants to (1) attend both workshops virtually via Zoom or in person beyond their paper-accepted workshop venue; (2) join the shared Slack platform that incorporates both workshops' attendees for active discussions and connections. By bridging the HCI and AI/ML communities, we aim to establish a shared platform for exploring alignment from interdisciplinary perspectives encompassing both academic and practical contexts.

Differences from Previous and Other Workshops. While existing workshops typically focus on specific aspects of alignment, such as human values or interactions, our workshop distinguishes them in three significant ways. First, we bridge interdisciplinary collaboration between HCI and AI researchers through practical initiatives and shared platforms. Notably, we host Bi-Align workshops at leading HCI (CHI) and ML (ICLR) conferences and establish a dedicated Slack platform to facilitate long-term communication and collaboration among participants beyond the workshop. Second, we offer a clarified definition and comprehensive roadmap for human-AI alignment based on a systematic literature review [6], which provides participants with a broader, more complete understanding of human-AI alignment. Third, we expand the traditional concept of static, one-directional "AI alignment" by introducing "bidirectional human-AI alignment," which highlights the dynamic, reciprocal nature of the alignment process. We further invite submissions to address less-explored challenges, such as integrating human values into AI design, understanding the dynamic impact of AI on humans, and incorporating these dynamics into alignment strategies.

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

- **Human Values in AI Alignment Design.** What are the core human values that AI must align with and how might they be integrated into designing AI that is aligned with human values?
- **Understanding and Evaluating Alignment.** To what extent does human understanding align with AI decision-making, and how can we effectively assess this alignment?
- **Alignment via User Experience and Interaction Design** How might we advance the end-user experience through the design of interactive systems that foster alignment?
- **Individual and Societal Impacts of AI Alignment.** How might we measure AI's impact on individuals and society, and refine alignment to incorporate these effects?
- **Dynamic Human-AI Co-evolution in Alignment.** How might we track the evolving nature of human-AI alignment and develop strategies to adapt to changes?

2 ORGANIZERS

This workshop is organized by a team of experienced researchers with diverse expertise in human-AI alignment, representing various stages of their careers and spanning a wide range of demographic and geographical backgrounds. Collectively, the organizing committee has made significant contributions to both HCI and AI across a variety of topics, including alignment and values [6, 7], AI explanation and sensemaking [9**?**], human-AI interaction [3, 11], responsible AI [], AI safety and alignment [], and NLP evaluation []. The organizers have a strong track record of successfully hosting workshops, tutorials, and panels on human-centered AI [], and human-AI interaction [], bridging the HCI and AI communities.

Hua Shen is a postdoctoral scholar at the University of Washington. Her research centers on bidirectional human-AI alignment, aiming to empower humans to interactively explain, evaluate, and collaborate with AI, while incorporating human feedback and values to improve AI systems. Her work is grounded in HCI and intersects with multiple AI fields, including Natural Language Processing, Speech Processing, and Computer Vision. She earned her Ph.D. from Pennsylvania State University and completed a postdoctoral fellowship at the University of Michigan.

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 from Pennsylvania State University, advised by Dr. John M. Carroll. She co-organized the CHI 2024 workshop on Computational UI.

Reshmi Ghosh is an Applied Scientist Lead for GenAI Safety in Microsoft's Responsible AI and Security team. She was the core architect in LLM Safety, designing M365 CoPilots, and has previously worked on integrating machine learning features to Excel, Word, and PowerPoint. She graduated with a Ph.D. from Carnegie Mellon University, focusing on data reconstruction using NLP methods for mitigating climate change. She is a research advisor for teams at MIT CSAIL, UMass Amherst, UCLA, and Oxford University.

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 previously earned his Ph.D. from the Human-Computer Interaction Institute at Carnegie Mellon University, specializing in the intersection of HCI, programming tools, sensemaking, intelligent user interfaces, and human-AI interaction. Michael organized the Sensemaking workshop at CHI 2024.

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 led the HCI Research team in the Microsoft Research's FUSE Labs and Snap Research. He received his Ph.D. degree in Media Arts and Sciences from MIT.

Sherry Tongshuang Wu is an Assistant Professor in the Human-Computer Interaction Institute at 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. Before joining CMU, Sherry received her Ph.D. degree from the University of Washington. Sherry organized the TRAIT workshop at CHI 2022, 2023 and TREW workshop at CHI 2024.

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 Socially Aware Natural Language Processing, aiming to better understand human communication in social context and build socially aware language technologies to support human-human and human-computer interaction. She received her Ph.D. degree in Language Technologies Institute at CMU.

Tanu Mitra is an Associate Professor in the Information School at the University of Washington. 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 harm. She co-founded the Responsibility in AI Systems and Experiences (RAISE) Center at UW. She received her PhD in Computer Science from Georgia Tech's School of Interactive Computing and her Masters in Computer Science from Texas A&M University.

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 organized multiple workshops that bridges HCI and AI/ML fields, including the first ICML AI&HCI workshp.

Marti A. Hearst is a Professor in the UC Berkeley School of Information and the Computer Science Division. She was Interim Dean and Head of School for the I School. Her research encompasses user interfaces with a focus on search, information visualization with a focus on text, computational linguistics, and educational technology. She co-founded the ACM Learning@Scale conference and is a former president of ACL, a CHI Academy member, the SIGIR Academy, an ACM Fellow, and an ACL Fellow. She received her PhD, MS, and BA degrees in Computer Science from UC Berkeley.

2.1 Co-Organizers at ICLR 2025 Bi-Align Workshop.

Martin Ziqiao Ma (Workflow Chair) 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. He received the Weinberg Cognitive Science Fellowship and Outstanding Paper Award, and organized several workshops at NLP/ML venues.

Antoine Bosselut is an Assistant Professor at EPFL in Lausanne, Switzerland, specializing in NLP and ML. His research focuses on building knowledge-enhanced language models that can reason and make inferences about the world. He was named to the Forbes 30 under 30 list in Science & Healthcare.

Furong Huang is an Associate Professor at the University of Maryland, specializing in trustworthy machine learning, AI for sequential decision-making, and high-dimensional statistics, She organized the NeurIPS competition, served as chair and organizer of NSF-Amazon Fairness in AI PI Meeting, Co-organizer of the NSF-IEEE workshop, and more.

Joyce Chai is a Professor at the University of Michigan. Her research interests span NLP and embodied AI to human-AI collaboration. She served on the executive board of NAACL and as Program Co-Chair for multiple conferences. She is a recipient of the NSF Career Award and multiple paper awards. She is a Fellow of ACL.

Dawn Song is a Professor 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, and more. She is an ACM Fellow and an IEEE Fellow.

3 PLANS TO PUBLISH WORKSHOP PROCEEDINGS

We plan to compile a comprehensive report summarizing the workshop's key discussions, presentations, and findings via open-access platforms, including ArXiv and our workshop website, allowing a broad audience to be informed about the content. In addition, we plan to collect the accepted workshop papers and curate an edited volume or a special journal issue in the journal such as ACM Transactions on Computer-Human Interaction (ToCHI) or ACM Transactions on Interactive Intelligent Systems (TIIS), or as a workshop proceedings at https://ceur-ws.org, where participants will be invited to contribute their work.

4 HYBRID FORMAT

We will host the workshop in a **hybrid format**, primarily in-person with an option for remote participation. All sessions will be live-streamed, with virtual breakout rooms available for remote attendees to join discussions. 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 pre-prints of accepted position papers.

5 ACCESSIBILITY

We strive to create an inclusive workshop environment for all participants, including those with cognitive, mental health and physical disabilities. We will highly encourage authors to make their position paper accessible. For accepted papers, we plan to offer guidance on improving document accessibility, e.g., alt-text for images and tables, ensuring that the navigation hierarchy is intelligible for screen-readers. During the workshop, we will request that all participants follow accessibility best practices, for example use of a microphone at all times and turning on captioning for presentations.

6 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 pre-prints 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.

7 WORKSHOP ACTIVITIES

Slot	Theme
09:00 – 09:15 (15min)	Welcome
09:15 – 10:15 (60min)	Keynote 1 by an invited speaker (Elizebeth Churchill, confirmed)
10:15 – 10:45 (30min)	Poster Session + Concurrent Coffee break
10:45 – 11:30 (45min)	Panel with experts that have diverse and well-balanced expertise
11:30 – 12:00 (30min)	Spotlight Paper sessions 1
12:00 – 13:30 (90min)	Lunch break
13:30 – 14:30 (60min)	Keynote 2 by an invited speaker (Brad Myers, confirmed)
14:30 – 15:00 (30min)	Poster Session + Concurrent Coffee break
15:00 – 15:30 (30min)	Spotlight Paper Session 2
15:30 – 17:00 (90min)	Group activity 2 *(60 min discussion + 30 min group result sharing)*
17:00 – 17:15 (15min)	Closing remarks
17:15	Dinner (optional)

Table 1. Tentative schedule for the proposed one-day CHI 2025 Bi-Align workshop.

We propose a single-day workshop, from 9:00 AM to 5:15 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, two keynotes and a brainstorming activity. Our overarching goal is to support participants to meaningfully connect with others in the blooming AI alignment community.

Keynote Talks. The morning and afternoon sessions will begin with keynotes. We have commitments from two leading experts from academia and industry with expertise in Human-Computer Interaction and AI: Elizebeth Churchill (Department Chair and Professor of Human Computer Interaction at MBZUAI, previous Director of User Experience at Google), and Brad Myers (Director and Professor of Human-Computer Interaction at Carnegie Mellon University). We've allotted 45 minutes for each speaker to give their talk followed by a 15 minute Q&A discussion.

Panel with Experts. The discussion panel will include experts with balanced perspectives from academia and industry who will touch on related workshop topics such as AI literacy, human-centered explanation, human-AI collaboration, the social impact of AI, the evolving role of humans in human-AI alignment, etc.

Spotlight Paper and Poster Sessions. 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 nominated by the program committees and decided by the organizers. It will be based on the paper's 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 7 minute lightning talks, concluded by a 10 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).

Group Activity. We plan to host a 90-minute group activity session in the afternoon for in-depth connection and discussion. The group activity includes a 60-minute smallgroup discussion followed by 30-minutes of insight sharing with another group. The groups will be formed using the "birds of afeather" format, which allows for individuals with shared interestes to engage in unstructured discussions about workshop topics. We will tailor group activities based on final participant numbers and interests. Our initial ideas include the following:

1. ***On-the-spot Paper Writing***: Participants will choose from predefined alignment topics and join corresponding groups. Each group will brainstorm a research idea or an imaginary paper on their chosen topic. Deliverables may include an abstract, teaser figures illustrating key concepts, or compelling use cases. Groups will then share their work with others for feedback.
2. ***Solution Ideation***: Groups will explore alignment solutions for specific use cases or contexts of practice, such as online education, news recommendation systems, or autonomous vehicles. They will brainstorm potential solutions, including UX interaction, explanation, or visualization, and iterate based on feedback from other groups. Outputs may include solution concepts or low-fidelity mockups for system or human study designs.
3. ***Concept Mapping***: Groups will collaboratively map out definitions and topics relevant to human-AI alignment, such as literacy, explanation, collaboration, and adaptability. The activity will focus on identifying relationships between these concepts, gaps in current research, and areas that should be prioritized in the future.

In addition to the above, we will encourage participants to join groups related to their paper topics so they can have dedicated discussions around a broader theme but in the context of their own work. For hybrid participation, activities will use collaborative virtual environments such as Google Documents and Miro boards.

Logistics. We will host the paper presentation spotlight recordings, posters, and artifacts of group work in Google Slides and on the website. After the workshop, we plan to convert them into Medium blog posts to share with a broad audience. We will utilize Zoom for live presentations and Slack for remote, asynchronous discussions and Q&A. Drawing from the success of previous TRAIT workshops, co-organized by our team member Tongshuang Sherry Wu, we plan to maintain a similar Slack setup, with dedicated channels for workshop sessions and individual threads for each accepted paper to facilitate focused questions and discussions. Keynote speakers, authors, and panelists will be encouraged to actively engage in these Slack threads. Additionally, we will collaborate with the CHI 2025 technical team to integrate streaming platforms, ensuring seamless access with captioning for improved accessibility and minimal platform transitions.

7.1 Confirmed Keynote Talks at the ICLR 2025 Bi-Align Workshop

Our participants can participate in both Bi-Align workshops at CHI and ICLR 2025 with access to all keynote talks. Therefore, we also list the keynote speakers of the Bi-Align at ICLR 2025 as supplementary events for this CHI workshop.

1. Been Kim (Senior Research Scientist, Google DeepMind). *Talk Topic*: Interpretability and Alignment.
2. Frauke Kreuter (Professor and Chair, LMU Munich and University of Maryland). *Talk Topic*: Dynamic Human Values, Preferences, and Social Norms.
3. Dan Bohus (Senior Principal Researcher, Microsoft Research). *Talk Topic*: Multimodal Situated Interaction.
4. Richard Ngo (Research Scientist, OpenAI). *Talk Topic*: AI Safety and Model Specification.

(5) Pavel Izmailov (Research Scientist at Anthropic / Assistant Professor at New York University,). *Talk Topic*: LLM Reasoning for Alignment and AI for Science.

(6) Hung-yi Lee (Associate Professor, National Taiwan University). *Talk Topic*: Alignment in Spoken Language Models.

8 POST-WORKSHOP PLANS

In addition to the post-workshop publishing plan outlined in Section 3, a key objective of the workshop is to foster ongoing community development among researchers and practitioners in this field. To sustain the momentum and facilitate ongoing dialogue, we plan to establish a platform for continued engagement and resource sharing. Options under consideration include a periodic email newsletter, a public GitHub repository, or a Slack/Discord channel. These next steps will be discussed with participants and organizers during the workshop.

9 CALL FOR PARTICIPATION

We invite authors from all disciplines to participate in the Bi-Align at CHI 2025 Workshop: "Bidirectional Human-AI Alignment (Bi-Align): Exploring Design, Interaction, Evaluation, and Critical Thinking of Alignment". This one-day event will bring together researchers from diverse fields, including HCI, ML, CV, NLP, and SE, fostering interdisciplinary collaboration on human-AI alignment. We welcome both academic and industry participants working on algorithm development, model applications, and human-centered AI approaches.

The workshop will focus on bidirectional alignment, integrating human perspectives into AI alignment research and practice. Our goal is to inspire discussions and novel insights on how AI can better align with human values, support critical thinking, and enhance collaboration between humans and AI.

We invite submissions of 3-6 page papers (excluding references) in CHI Extended Abstract format. Submissions may include position papers, syntheses of ongoing research, or descriptions of interactive systems. Each paper will be lightly reviewed by at least two program committee members, with selection based on quality and relevance.

Papers can focus on potentiall topics, but not limited to:

- integrating human values into AI design
- Understanding and Evaluating Alignment.
- fostering critical thinking about AI;
- designing productive human-AI interactions;
- enhancing UX for AI systems
- exploring the societal impact of AI

Submission Requirements. Visit our website for submission details: https://bialign-workshop.github.io/ and submit via https://openreview.net/. We look forward to your contributions and to advancing the field of human-AI alignment through collaborative dialogue.

10 EXPECTED SIZE OF ATTENDANCE

In order to facilitate meaningful, in-depth conversation, we have tailored this workshop for a group of **30-50** participants, including 30-40 in-person and 10-20 remote participants. If interested participants exceed this number after the initial call of the workshop, we may adjust the workshop structure to accommodate a slightly higher number of participants.

REFERENCES

[1] Micah Carroll, Davis Foote, Anand Siththaranjan, Stuart Russell, and Anca Dragan. 2024. AI Alignment with Changing and Influenceable Reward Functions. *arXiv:2405.17713* (2024).

[2] Nitesh Goyal, Minsuk Chang, and Michael Terry. 2024. 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*. 1–6.

[3] Qianou Ma, Hua Shen, Kenneth Koedinger, and Tongshuang Wu. 2024. How to Teach Programming in the AI Era? Using LLMs as a Teachable Agent for Debugging. *25th International Conference on Artificial Intelligence in Education (AIED 2024)* (2024).

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

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

[6] Hua Shen, Tiffany Knearem, Reshmi Ghosh, Kenan Alkiek, Kundan Krishna, Yachuan Liu, Ziqiao Ma, Savvas Petridis, Yi-Hao Peng, Li Qiwei, et al. 2024. Towards Bidirectional Human-AI Alignment: A Systematic Review for Clarifications, Framework, and Future Directions. *arXiv preprint arXiv:2406.09264* (2024).

[7] Hua Shen, Tiffany Knearem, Reshmi Ghosh, Yu-Ju Yang, Tanushree Mitra, and Yun Huang. 2024. ValueCompass: A Framework of Fundamental Values for Human-AI Alignment. *arXiv preprint arXiv:2409.09586* (2024).

[8] Michael Terry, Chinmay Kulkarni, Martin Wattenberg, Lucas Dixon, and Meredith Ringel Morris. 2023. AI Alignment in the Design of Interactive AI: Specification Alignment, Process Alignment, and Evaluation Support. *arXiv:2311.00710* (2023).

[9] **Hua Shen**, Chieh-Yang Huang, Tongshuang Wu, and Ting-Hao 'Kenneth' Huang. 2023. ConvXAI: Delivering Heterogeneous AI Explanations via Conversations to Support Human-AI Scientific Writing.. In *The 26th ACM Conference On Computer-Supported Cooperative Work And Social Computing - Demo (CSCW '23 Demo)* ***Best Demo***.

[10] Wikipedia. 2024. AI alignment — Wikipedia, The Free Encyclopedia. http://en.wikipedia.org/w/index.php?title=AI%20alignment&oldid=1220304776. [Online; accessed 05-May-2024].

[11] Tongshuang Wu, Michael Terry, and Carrie Jun Cai. 2022. Ai chains: Transparent and controllable human-ai interaction by chaining large language model prompts. In *Proceedings of the 2022 CHI conference on human factors in computing systems*. 1–22.