Tutorial: Human-AI Alignment: Foundations, Methods, Practice, and Challenges

Hua Shen* [*]
NYU Shanghai,
New York University
huashen@nyu.edu

Mitchell Gordon*
OpenAI,
MIT
mlgordon@mit.edu

Adam Tauman Kalai
OpenAI
adam@kal.ai

Abstract

The rapid progress of general-purpose AI systems has created an urgent need to align these technologies with human values, ethics, and societal goals. While traditional approaches treat alignment as a static, one-way process, this tutorial frames it as a dynamic, bidirectional relationship in which humans and AI systems continuously adapt to one another. We introduce a structured framework for Human-AI Alignment and systematically examine how to empower human agency throughout the alignment process. The tutorial is organized around three core areas: Foundations (what values should AI align with?), Methods (how can we empower humans in alignment across system stages?), and Practice (what sociotechnical impacts result from AI deployment?). The session concludes with a multidisciplinary panel featuring four leading experts discussing emerging Challenges and future directions in alignment research. This tutorial equips participants with essential conceptual foundations, practical methodologies, and critical perspectives on the evolving alignment landscape. All materials, including slides, coding resources, and recordings, will be publicly available through our tutorial website[2].

1 Description

The rapid advancement of general-purpose AI has created an urgent need to align these systems with human values, ethical principles, and societal goals. This challenge, known as AI alignment [1], is critical for ensuring that AI systems function effectively while minimizing harm and maximizing societal benefits. Traditionally, AI alignment has been conceptualized as a static, unidirectional process aimed at shaping AI systems to achieve desired outcomes and prevent adverse consequences [2]. However, this one-way approach is insufficient, as AI systems increasingly interact with humans in dynamic, unpredictable ways, creating feedback loops that influence both AI behavior and human responses [3]. This **evolving interaction requires a fundamental shift** toward recognizing the bidirectional and adaptive nature of human-AI relationships [4].

While previous alignment tutorials have primarily treated AI alignment as a static process of fitting AI to human and institutional expectations, this tutorial frames alignment as a **continuous, evolving engagement between humans and AI**. To clarify the dynamic roles of humans and AI, we introduce a conceptual framework for Human-AI Alignment (Figure 1) and systematically explain **how humans can be empowered throughout the stages of the alignment process**. Specifically, the tutorial addresses three central questions: (i) **Foundations** — what do humans expect AI to align with? (ii) **Methods** — how can humans be empowered in building aligned AI? and (iii) **Practice** — what sociotechnical impacts does AI have on humans and society? To stimulate debate and inform future research directions, the tutorial concludes with a comprehensive discussion on (iv) **Challenges**,

[*] Equal Contributions

[2] Tutorial Website: https://hai-alignment-course.github.io/tutorial/

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

incorporating the perspectives of three speakers and four panelists from diverse backgrounds on emerging topics and open problems in Human-AI Alignment.

Goals. This tutorial aims to benefit audience through four main objectives: (1) **Comprehensive Overview**: Provide a systematic perspective on Human-AI Alignment, emphasizing human engagement throughout the alignment process. (2) **Knowledge and Understanding**: Offer comprehensive knowledge on human values, alignment techniques, and the societal impacts of AI systems. (3) **Practical Skills**: Equip participants with interactive coding notebooks and hands-on exercises, enabling them to practice actionable tools and implement basic alignment strategies in diverse AI systems. (4) **Facilitate Discussion**: Foster critical dialogue on future challenges, open problems, and emerging opportunities in alignment research, offering research inspiration for participants' future work.

Importance and Impacts. Given the existing challenges that alignment challenges in existing frameworks are ill-equipped to address, there is an increasing need for professionals who understand the full scope of Human-AI Alignment, encompassing both technical foundations and sociotechnical implications. This tutorial addresses this gap by preparing participants to contribute meaningfully to alignment research, policy discussions, and implementation efforts. By offering a balanced combination of conceptual frameworks, technical methods, and critical debate, the tutorial ensures participants engage with the current state of the field rather than treating alignment as a resolved issue. The interactive panel component further cultivates the analytical skills and perspectives necessary to navigate and advance this rapidly evolving area.

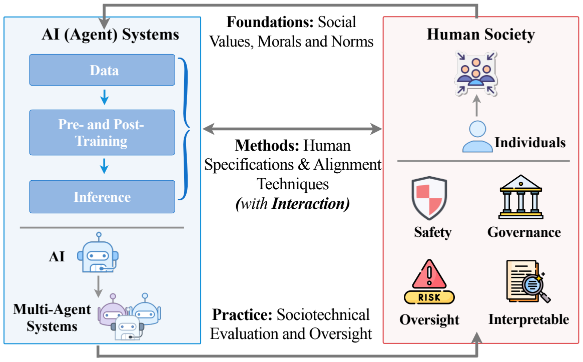
AI (Agent) Systems
Data
Pre- and Post-Training
Inference
AI
Multi-Agent Systems
Foundations: Social Values, Morals and Norms
Methods: Human Specifications & Alignment Techniques (with Interaction)
Practice: Sociotechnical Evaluation and Oversight
Human Society
Individuals
Safety
Governance
RISK
Oversight
Interpretable

Figure 1: A conceptual framework for Human-AI Alignment.

2 Related Tutorials and Workshops

Alignment research has gained significant momentum over the past three years, with an increasing number of workshops exploring related themes. Notable examples include: ICLR 2025 Workshop on Bidirectional Human-AI Alignment [link]; NeurIPS 2024 Workshop on Pluralistic Alignment [link]; ICML 2025 Workshop on Models of Human Feedback for AI Alignment [link]; ICLR 2025 Workshop on Representational Alignment [link]. In comparison, we only identified one tutorial in previous venues, featured in NeurIPS 2024: the tutorial on "Cross-disciplinary insights into alignment in humans and machines" [slide and video]. While this tutorial introduced cross-disciplinary perspectives and emphasized Institution-Compatible Alignment as the primary optimization goal, there is still a lack of systematic, in-depth examination of humans' roles in alignment and concrete approaches for empowering humans throughout the alignment process.

Distinctive Contributions of This Tutorial. This tutorial differs from previous offerings in three key aspects. **First**, we broaden the traditional, static concept of AI alignment by introducing a conceptual framework for Human-AI Alignment, capturing the dynamic and evolving nature of alignment. **Second**, we provide structured, actionable roadmaps for empowering humans multiple stages of the alignment process, combining theoretical foundations with practical tools. **Third**, we convene panelists from varied disciplines and backgrounds to collectively discuss emerging challenges and future directions in Human-AI Alignment. Overall, this tutorial offers a comprehensive examination of the interactive, reciprocal process of human-AI collaboration in achieving dynamic, bidirectional alignment, addressing a critical gap in current educational resources.

3 Outline of Tutorial Content

In the following, we provide an outline for the 2.5-hour tutorial, supplied with important References that will be covered in each section.

I. Introduction (15 min)
Why do we need to align AI with humans, and why should humans be empowered in the alignment process? This section introduces the motivations behind AI alignment as a requirement for AI safety and necessity of human engagement. We present a conceptual framework for Human-AI Alignment (Figure 1) and outline the structure of the tutorial guided by this framework.

- Overview of the tutorial;
- Basics and recent progress of human-AI alignment;
- Importance of alignment for AI safety;
- Importance of empowering humans in the alignment process;
- References: AI Risks and Safety: [5, 6, 7, 8], Alignment: [9, 10, 4]

II. Foundations: Pluralistic Values, Morals and Norms (25min)
Which values do humans expect AI systems to align with? The section introduces taxonomies, representative value theories, datasets and methods for pluralistic value alignment.

- A taxonomy of AI values and morals;
- Representative value and theories and datasets;
- Value evaluation and validation approaches;
- References: Overview and taxonomy of alignment goals: [11, 12, 13], Value and moral theories: [14, 15, 16, 17, 18, 19], Datasets and evaluation: [20, 21, 22, 23, 24].

III. Methods: Human Specifications and Alignment Techniques (30min)
How can we empower humans in building aligned AI systems? This section encompasses training, optimization, evaluation, multi-agent systems, and technical implementation.

- Human specifications and data collection strategies;
- Alignment techniques at pre-training, post-training, and inference stages;
- Interactive, customized alignment techniques;
- Evaluation protocols and benchmarks;
- References: Specification methods: [25, 26, 27**?**], Alignment techniques: [28, 29, 30, 31, 32, 33], Interactive Alignment: [34, 35, 36], Evaluation and benchmarks: [37, 38, 39].

IV. Practice: Sociotechnical Evaluation and Oversight (25min) + Q&A (10min)
What sociotechnical impacts does AI have on humans and society? This section explores the broader societal consequences of AI alignment, and the dynamic influence of AI on humans and institutions.

- Alignment for safety and trustworthiness;
- Interpretability, controllability, and oversight;
- LLM simulations and sociotechnical impacts;
- References: Safety and trust: [40, 41], Interpretability and controllability: [42, 43, 44], Simulations and social impacts: [45, 46, 47, 48, 49]

V. Challenges: Emerging Topics and Future Directions (15min)
What are the emerging technical and ethical challenges in achieving human-AI alignment?

- Dynamic and Evolving Alignment
- Tradeoff between Safety and Performance
- Deceptive Alignment and Alignment Faking
- Alignment in Agentic AI and Multi-agent Systems
- References: Dynamic and evolving alignment: [3, 3], Safety-performance trade-off: [50, 51], Deceptive alignment: [52, 53], Multi-agent alignment: [54]

VI. Panel: Challenges and Prospects (30min)
Four invited panelists from diverse sectors and global backgrounds will share perspectives on challenges, open problems, and future directions in alignment research, providing participants with a multidimensional perspective. Hua Shen will be the moderator. **Confirmed** panelists include:

- **Yoshua Bengio** (Professor, Mila - Quebec AI Institute, `yoshua.bengio@mila.quebec`)
  Bio: Prof. Bengio is the 2018 A.M. Turing Award laureate and one of the most cited and h-index researchers in AI. He is full professor at Université de Montréal, and Founder and Scientific Advisor of Mila – Quebec AI Institute. He received numerous awards, such

as the prestigious Killam Prize and Herzberg Gold medal in Canada, and member of the UN's Scientific Advisory Board for Independent Advice on Breakthroughs in Science and Technology. Prof. Bengio was named in 2024 one of TIME's magazine 100 most influtential people in the world. He is internationally recognized for his contributions to deep learning and AI safety. He actively engages in alignment and responsible development of AI, including chairing the International Scientific Report on the Safety of Advanced AI.

- **Dawn Song** (Professor, UC Berkeley, dawnsong@cs.berkeley.edu)

  Bio: Prof. Song is a Professor in the Department of EECS at UC Berkeley. She is an expert at the intersection of AI, security, and privacy, addressing the technical and ethical challenges of AI systems. As an ACM and IEEE Fellow, her research includes developing secure, trustworthy AI systems — contributing directly to alignment concerns in safety-critical and privacy-sensitive applications. 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 ranked the most cited scholar in computer security (AMiner Award).

- **Eric Gilbert** (Professor, University of Michigan, http://eegilbert.org/)

  Bio: Eric Gilbert is a Professor at the University of Michigan, where he directs the comp.social lab. A leading sociotechnologist, his research centers on the design, development, and study of social computing systems. He was a Visiting Scholar with the Rebooting Social Media initiative at Harvard's Berkman Klein Center, having previously served on the faculty at Georgia Tech. His work is highly recognized, having received the NSF CAREER award, the National Academy of Sciences Kavli Fellowship, the 2024 ICWSM Test of Time award, and multiple best paper awards. He is also a Distinguished Member of the ACM.

- **Monojit Choudhury** (Professor, MBZUAI, monojit.choudhury@mbzuai.ac.ae)

  Bio: Prof. Choudhury is a full professor of Natural Language Processing at MBZUAI. His research interests center around the convergence of Language Technology and Society, exploring pivotal inquiries such as the learning and (mis)representation of linguistic and cultural diversity by foundation models, contributing to value alignment and representational fairness in global AI deployments. His research delves into the impact of representational disparities on present and future of technology use, and their impact on linguistic and cultural dynamics in real world. He is the recipient of various awards, including the Young Scientist Award by Indian Science Congress Association. He also served as multiple roles, such as Associate editor of IEEE Trans. on Pattern Analysis and Machine Intelligence.

- **Hannah Kirk** (Research Scientist, UK AI Security Institute, hannah@thekirks.co.uk)

  Bio: Hannah Kirk is the Research Scientist at UK AI Security Institute and Ph.D. at the University of Oxford. Her research centres on human-and-model-in-the-loop feedback and data-centric alignment of AI, particularly emphasizing the societal impact of AI systems as we scale across model capabilities, domains and human populations. Her body of published work spans computational linguistics, economics, ethics and sociology, addressing a broad range of issues such as alignment, bias, fairness and hate speech from a multidisciplinary perspective. She has been invited as a keynote talk at various venues including the NeurIPS 2024 Pluralistic Workshop.

4 Speakers, Panel, and Diversity

4.1 Speakers

All three speakers have **confirmed** to attend the tutorial and each will deliver at least one 30-minute session part. Their detailed biographical information and areas of expertise are presented below.

Hua Shen (Assistant Professor at NYU Shanghai, New York University, huashen@nyu.edu)

Bio: Hua Shen is an Assistant Professor of Computer Science at New York University Shanghai and a Postdoctoral Scholar at the University of Washington, who initiated and leads bidirectional human-AI alignment research. Her research spans HCI, NLP, speech processing, and computer vision. Her works on human-AI interaction and alignment have received multiple awards (e.g., AIED'24 Best

Paper, CSCW'23 Best Demo, Google Scholarship). She is a recipient of 2023 Rising Star of Data Science. She was invited to talk at leading academic and industrial institutions, including CMU, UW, UIUC, UMich, Princeton University, Google DeepMind, Microsoft, and more. Dr. Shen serves in leading roles at top-tier conferences (CHI Associate Chair, ACL/EMNLP program committees) and founded the Bidirectional Human-AI Alignment workshops at ICLR 2025 and CHI 2025. **Talk Recordings**: ICLR 2025 BiAlign Workshop Opening Remarks., UW Data Science Seminar.

Mitchell Gordon (Assistant Professor at MIT, Researcher at OpenAI, `mlgordon@mit.edu`)

Bio: Mitchell Gordon is an Assistant Professor of Computer Science at MIT CSAIL and CCES, and a Researcher at OpenAI. He leads research at the intersection of human-computer interaction and machine learning, with a focus on designing interactive systems, evaluation approaches, and pluralistic values for human-AI alignment. His work has received multiple recognitions, including a CHI Best Paper Award, CSCW Best Paper Award, Apple PhD Fellowship in AI/ML, Oral Presentation at NeurIPS, and more. Dr. Gordon organized the NeurIPS 2024 Pluralistic Alignment Workshop and has been invited to speak at multiple AI conferences and academic institutions. He actively serves in leadership roles across top-tier venues, including NeurIPS, CHI and UIST Area Chairs. He completed his Ph.D. at Stanford advised by Profs. Michael Bernstein and James Landay, and a Postdoc Scholar at University of Washington advised by Prof. Yejin Choi. **Talk Recording**: Stanford Seminar.

Adam Tauman Kalai (Research Scientist, OpenAI, `adam@kal.ai`)

Bio: Dr. Kalai is an American computer scientist who specializes in AI and works at OpenAI, specializing in AI Safety, Fairness, Ethics, and alignment of advanced AI systems. His research spans algorithms, fairness, machine learning theory, game theory, crowdsourcing, and the social implications of generative models. Prior to OpenAI, he was a Senior Principal Researcher at Microsoft Research New England from 2008 to 2023. Dr. Kalai's work has received broad recognition for advancing technical methods and conceptual understanding in AI ethics and alignment. He has co-chaired leading conferences such as COLT (Conference on Learning Theory), HCOMP (Conference on Human Computation), and NEML. His work has been recognized with numerous honors, including several best paper awards, an NSF CAREER Award, an Alfred P. Sloan Fellowship, and most notably the Majulook Prize. **Talk Recording**: Keynote at 2024 ACM Conf. on Economics and Computation.

4.2 Diversity of Speakers and Panelists

The speakers and panelists for this tutorial represent a broad spectrum of demographic backgrounds, professional experiences, and research perspectives. We actively promote diversity along multiple axes, including **gender, race, geographical location, institutional affiliation, academic seniority, and research specialty**. Among the 3 speakers and 4 panelists, 3 identify as female and 4 as male. The group includes 5 academic researchers — comprising 3 full professors and 2 assistant professors — alongside 1 researcher from a leading AI industry lab (OpenAI) and 1 researcher from the UK government's AI Security Institute. Geographically, participants are based in USA, Canada, Asia, Europe, and the United Arab Emirates, reflecting international perspectives on human-AI alignment.

In terms of expertise, the speakers and panelists bring complementary strengths across AI safety and alignment, human-AI interaction, algorithmic fairness, interpretability, language and multimodal models, AI ethics, policy, and societal impacts. This diversity ensures that the tutorial and panel discussions reflect a range of technical, methodological, and cultural viewpoints, fostering balanced and inclusive dialogue on emerging challenges and opportunities in Human-AI Alignment.

5 Tutorial Website and Resource Access

This tutorial will provide lasting value to attendees through multiple accessible channels:

Tutorial Website: we will maintain a dedicated tutorial website[3] to host all materials, including presentation slides, hands-on coding notebooks, demonstration resources, recommended papers and reading. Attendees will be able to reproduce key demonstrations and exercises using the provided interactive notebooks.

[3] Tutorial Website: https://hai-alignment-course.github.io/tutorial/

Video Archive and Audio-Accessible Format:: All presentations will be recorded and made available via the NeurIPS archive and YouTube, with audio-accessible content. Presentations will remain comprehensible in audio-only format, ensuring accessibility for a broader audience.

Hybrid and Asynchronous Access: The tutorial will support in-person and remote participation, with all materials accessible online after the event for flexible, asynchronous engagement.

References

[1] 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.

[2] 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.

[3] Micah Carroll, Davis Foote, Anand Siththaranjan, Stuart Russell, and Anca Dragan. Ai alignment with changing and influenceable reward functions. In Proceedings of the 41st International Conference on Machine Learning, ICML'24. JMLR.org, 2024.

[4] Hua Shen, Tiffany Knearem, Reshmi Ghosh, Kenan Alkiek, Kundan Krishna, Yachuan Liu, Ziqiao Ma, Savvas Petridis, Yi-Hao Peng, Li Qiwei, et al. Towards bidirectional human-ai alignment: A systematic review for clarifications, framework, and future directions. arXiv preprint arXiv:2406.09264, 2024.

[5] Yoshua Bengio, Geoffrey Hinton, Andrew Yao, Dawn Song, Pieter Abbeel, Trevor Darrell, Yuval Noah Harari, Ya-Qin Zhang, Lan Xue, Shai Shalev-Shwartz, et al. Managing extreme ai risks amid rapid progress. Science, page eadn0117, 2024.

[6] Yoshua Bengio, Sören Mindermann, Daniel Privitera, Tamay Besiroglu, Rishi Bommasani, Stephen Casper, Yejin Choi, Philip Fox, Ben Garfinkel, Danielle Goldfarb, et al. International ai safety report. arXiv preprint arXiv:2501.17805, 2025.

[7] Laura Weidinger, John Mellor, Maribeth Rauh, Conor Griffin, Jonathan Uesato, Po-Sen Huang, Myra Cheng, Mia Glaese, Borja Balle, Atoosa Kasirzadeh, et al. Ethical and social risks of harm from language models. arXiv preprint arXiv:2112.04359, 2021.

[8] 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, pages 23082–23104. PMLR, 2024.

[9] Usman Anwar, Abulhair Saparov, Javier Rando, Daniel Paleka, Miles Turpin, Peter Hase, Ekdeep Singh Lubana, Erik Jenner, Stephen Casper, Oliver Sourbut, et al. Foundational challenges in assuring alignment and safety of large language models. arXiv:2404.09932, 2024.

[10] Stephen Casper, Xander Davies, Claudia Shi, Thomas Krendl Gilbert, Jérémy Scheurer, Javier Rando, Rachel Freedman, Tomasz Korbak, David Lindner, Pedro Freire, et al. Open problems and fundamental limitations of reinforcement learning from human feedback. arXiv preprint arXiv:2307.15217, 2023.

[11] Iason Gabriel. Artificial intelligence, values, and alignment. Minds and machines, 30(3):411–437, 2020.

[12] Haoran Ye, Jing Jin, Yuhang Xie, Xin Zhang, and Guojie Song. Large language model psychometrics: A systematic review of evaluation, validation, and enhancement. arXiv preprint arXiv:2505.08245, 2025.

[13] 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 preprint arXiv:2402.05070, 2024.

[14] Shalom H Schwartz, Jan Cieciuch, Michele Vecchione, Eldad Davidov, Ronald Fischer, Constanze Beierlein, Alice Ramos, Markku Verkasalo, Jan-Erik Lönnqvist, Kursad Demirutku, et al. Refining the theory of basic individual values. Journal of personality and social psychology, 103(4):663, 2012.

[15] Samuel Cahyawijaya, Delong Chen, Yejin Bang, Leila Khalatbari, Bryan Wilie, Ziwei Ji, Etsuko Ishii, and Pascale Fung. High-dimension human value representation in large language models. arXiv preprint arXiv:2404.07900, 2024.

[16] Dorit Hadar-Shoval, Kfir Asraf, Yonathan Mizrachi, Yuval Haber, and Zohar Elyoseph. Assessing the alignment of large language models with human values for mental health integration: cross-sectional study using schwartz's theory of basic values. JMIR Mental Health, 11:e55988, 2024.

[17] Yu Ying Chiu, Liwei Jiang, and Yejin Choi. Dailydilemmas: Revealing value preferences of llms with quandaries of daily life. arXiv preprint arXiv:2410.02683, 2024.

[18] Julia Kharchenko, Tanya Roosta, Aman Chadha, and Chirag Shah. How well do llms represent values across cultures? empirical analysis of llm responses based on hofstede cultural dimensions. arXiv preprint arXiv:2406.14805, 2024.

[19] Han Jiang, Xiaoyuan Yi, Zhihua Wei, Ziang Xiao, Shu Wang, and Xing Xie. Raising the bar: Investigating the values of large language models via generative evolving testing. arXiv preprint arXiv:2406.14230, 2024.

[20] 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.

[21] Yuanyi Ren, Haoran Ye, Hanjun Fang, Xin Zhang, and Guojie Song. Valuebench: Towards comprehensively evaluating value orientations and understanding of large language models. arXiv preprint arXiv:2406.04214, 2024.

[22] Haoran Ye, Yuhang Xie, Yuanyi Ren, Hanjun Fang, Xin Zhang, and Guojie Song. Measuring human and ai values based on generative psychometrics with large language models. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 39, pages 26400–26408, 2025.

[23] 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.

[24] Liwei Jiang, Taylor Sorensen, Sydney Levine, and Yejin Choi. Can language models reason about individualistic human values and preferences? ACL, 2025.

[25] 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, pages 853–868, 2024.

[26] 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, pages 1487–1505, 2023.

[27] Mitchell L Gordon, Michelle S Lam, Joon Sung Park, Kayur Patel, Jeff Hancock, Tatsunori Hashimoto, and Michael S Bernstein. Jury learning: Integrating dissenting voices into machine learning models. In Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems, pages 1–19, 2022.

[28] 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.

[29] 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.

[30] 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.

[31] 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, pages 10383–10405, 2023.

[32] Mirac Suzgun and Adam Tauman Kalai. Meta-prompting: Enhancing language models with task-agnostic scaffolding. arXiv preprint arXiv:2401.12954, 2024.

[33] Eric Zelikman, Eliana Lorch, Lester Mackey, and Adam Tauman Kalai. Self-taught optimizer (stop): Recursively self-improving code generation. In First Conference on Language Modeling, 2024.

[34] 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, pages 5994–6017, 2022.

[35] 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.

[36] 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.

[37] 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.

[38] 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.

[39] 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.

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

[41] Silen Naihin, David Atkinson, Marc Green, Merwane Hamadi, Craig Swift, Douglas Schonholtz, Adam Tauman Kalai, and David Bau. Testing language model agents safely in the wild. arXiv preprint arXiv:2311.10538, 2023.

[42] 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), pages 7283–7298, 2024.

[43] Shangbin Feng, Taylor Sorensen, Yuhan Liu, Jillian Fisher, Chan Young Park, Yejin Choi, and Yulia Tsvetkov. Modular pluralism: Pluralistic alignment via multi-llm collaboration. arXiv preprint arXiv:2406.15951, 2024.

[44] Hua Shen, Chieh-Yang Huang, Tongshuang Wu, and Ting-Hao Kenneth Huang. Convxai: Delivering heterogeneous ai explanations via conversations to support human-ai scientific writing. In Companion Publication of the 2023 Conference on Computer Supported Cooperative Work and Social Computing, pages 384–387, 2023.

[45] Gati V Aher, Rosa I Arriaga, and Adam Tauman Kalai. Using large language models to simulate multiple humans and replicate human subject studies. In International Conference on Machine Learning, pages 337–371. PMLR, 2023.

[46] Chance Jiajie Li, Jiayi Wu, Zhenze Mo, Ao Qu, Yuhan Tang, Kaiya Ivy Zhao, Yulu Gan, Jie Fan, Jiangbo Yu, Jinhua Zhao, et al. Position: Simulating society requires simulating thought. arXiv e-prints, pages arXiv–2506, 2025.

[47] Sayash Kapoor, Rishi Bommasani, Kevin Klyman, Shayne Longpre, Ashwin Ramaswami, Peter Cihon, Aspen Hopkins, Kevin Bankston, Stella Biderman, Miranda Bogen, et al. On the societal impact of open foundation models. arXiv preprint arXiv:2403.07918, 2024.

[48] 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, pages 1–22, 2023.

[49] Joon Sung Park, Carolyn Q Zou, Aaron Shaw, Benjamin Mako Hill, Carrie Cai, Meredith Ringel Morris, Robb Willer, Percy Liang, and Michael S Bernstein. Generative agent simulations of 1,000 people. arXiv preprint arXiv:2411.10109, 2024.

[50] Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, et al. Training a helpful and harmless assistant with reinforcement learning from human feedback. arXiv preprint arXiv:2204.05862, 2022.

[51] Robert Kirk, Ishita Mediratta, Christoforos Nalmpantis, Jelena Luketina, Eric Hambro, Edward Grefenstette, and Roberta Raileanu. Understanding the effects of rlhf on llm generalisation and diversity. arXiv preprint arXiv:2310.06452, 2023.

[52] Hua Shen, Nicholas Clark, and Tanushree Mitra. Mind the value-action gap: Do llms act in alignment with their values? arXiv preprint arXiv:2501.15463, 2025.

[53] Ryan Greenblatt, Carson Denison, Benjamin Wright, Fabien Roger, Monte MacDiarmid, Sam Marks, Johannes Treutlein, Tim Belonax, Jack Chen, David Duvenaud, et al. Alignment faking in large language models. arXiv preprint arXiv:2412.14093, 2024.

[54] Atrisha Sarkar, Andrei Ioan Muresanu, Carter Blair, Aaryam Sharma, Rakshit S Trivedi, and Gillian K Hadfield. Normative modules: A generative agent architecture for learning norms that supports multi-agent cooperation. arXiv preprint arXiv:2405.19328, 2024.