Foundations of Human-AI Interactions: Theory, Algorithms, Practice

ICML 2026 Workshop Proposal

Website: https://human-ai-interaction-workshop.github.io/

Abstract

[RC: sorry this is not good i need to rewrite this]Machine Learning systems today, particularly Generative Models and Large Language Models, have transitioned from static, one-shot predictors to active agents in multi-round, adaptive collaboration. Despite major advances in their usability and capabilities, current literature lacks a unified, foundational and principled understanding on Human-AI Interactions, such as exploring the key drivers of successful collaborations and deciding when is the model's response optimal for decision-making. Our workshop aims to build the foundations of HAI by bridging the different perspectives from existing fields, such as machine learning, statistics, optimization and decision-theory. By fostering an environment with holistic perspectives, we seek to motivate more emphasis of human aspects in today's theory, practice and algorithms for AI.

Motivation. Collaboration has always been central to knowledge creation and decision making. From negotiation and collective problem solving to scientific discovery, progress often emerges through interaction rather than isolated reasoning. As today's machine learning systems such as Large Language Models (LLMs) [1, 8, 17, 24, 26] and Diffusion Models [11, 22, 27, 28] are increasingly generative, capable and versatile, they become active participants in our collaborations and task. Through multi-turn, conversational exchanges, we as humans often use AI to explore ideas, verify reasoning, navigate uncertainty, and support decisions across domains. Hence, Human–AI Interaction (HAI) is taking forms of collaborative, multi-agent behavior, often involving asymmetric capabilities, responsibilities, and incentives between humans and machines.

Despite its growing importance, much of our understanding of HAI has lacked a rigorous, foundational understanding. Questions such as, "when is Human-AI Collaboration actually helpful?" [9], "What is optimal collaboration when multiple parties hold different kind of information?" [7], or "What is the appropriate notion of interpretability in today's multi-turn, collaborative, agentic systems?" [13] are broadly raised, but have yet to arrive at a clear answer formed by our research community. As AI systems become more interactive, adaptive, and deeply integrated into decision-making workflows, there is an increasing need for complementary foundations that identify general principles capable of guiding algorithm design, providing guarantees, and supporting systematic evaluation. In parallel, classical machine learning theory and optimization have traditionally modeled humans as sources of labels or feedback, leaving open the challenge of formally capturing humans as adaptive collaborators embedded within learning and inference loops.

Scope. The goal of this workshop is to establish a foundational science of HAI, bridging perspectives from well-established areas of research (e.g. machine learning, statistics, decision-theory and economics) and identifying principled theory, algorithms and practice that govern effective collaboration between humans and AI systems. We aim to bring together researchers working across theory, algorithms and practice to develop a shared language and tools for studying human-AI interaction. We invite contributions along three tightly connected dimensions:

1. Theory: foundational principles and formal models of human-AI interaction, including statistical learning theory for interaction, uncertainty quantification and decision making with humans in the loop [18], value of information, game-theoretic and sequential models of collaboration [5–7], and theoretical characterizations of human responses and behavior [9].
2. Algorithms: Algorithmic methods that operationalize these principles such as learning with human feedback [19], interactive and conversational systems [3, 16], uncertainty aware inference, interpretability and explanation as interaction mechanism [13], and evaluation metrics that go beyond standalone model accuracy to capture collaborative performance.
3. Practice: real-world deployments and use cases where interaction is essential, including healthcare and medicine [4, 15], scientific discovery [23], recommendation systems [12], entertainment [25], embodied agents and robotics [21], personalization, preference elicitation and scenarios involving partial information or underspecified prompts [2, 10, 14, 20]. We invite practioners to share assumptions and examples of real-world human-AI interactions to motivate better algorithms.

Workshop Contributions.We summarize the key contributions of our workshop as follows:

1. Exchange of Perspective from Holistic, Established Fields. We aim to bring together research across theory and practice. Through such exchange, we aim to foster and build a community that develops methods with a stronger emphasis on HAI.
2. Highlight Diverse Industry and Academic Perspectives. Featuring speakers from leading companies (Google DeepMind, Microsoft) and 3 distinct academic institutions, we bridge Featuring speakers from leading companies (Apple, Google) and 5 distinct academic institutions, we bridge the gap between various industry and academia settings. Each speaker represents an established field (i.e. uncertainty quantification, interpretability, optimization, decision-theory etc.) within health time-series research, ensuring broad relevance for attendees and fostering cross-disciplinary dialogue.

3. Meaningfully Engage the Larger Community. Through dual poster sessions, spotlight talks, and interactive panels, we facilitate idea exchange and collaboration among researchers. Dedicated Q&A sessions and networking breaks further provide opportunities to connect with speakers and senior community members. Our long-term goal is to move the research community for HAI forward and establish HAI as an impactful field of research.

History of Workshop and Previous Workshops. We are the first to organize the workshop on this topic. To the best of our knowledge, there exist two workshops that are related but distinctly differ from our workshop: First, *NeurIPS 2025 Workshop on Multi-Turn Interactions in Large Language Models* explores the methodology, evaluation and applications multi-turn usage of LLMs. While performing tasks with LLMs is often a primary example, we argue the formal definition of HAI precedes the choice of model itself, where HAI also exists in medical diagnosis, scientific discovery and creative arts [], and more. In contrast, our workshop poses no restrictions on the AI model and emphasizes more on building a strong foundation for understanding where HAI is meaningful. Second, *MICCAI Workshop on Human-AI Collaboration (HAIC)* explores how medical professionals may benefit from HAI. While the theme is aligned, it aims to explore practices in the medical domain, such as AI-assisted radiology report generation, rather than the mathematical foundation of human-AI collaboration. Overall, we seek to build a research community that establish a principled understanding of HAI.

Invited Speakers (All Confirmed).

- Dr. Been Kim, Senior Staff Research Scientist at Google DeepMind
  Research Areas: Interpretability
- Dr. Jessican Hullman, Ginni Rometty Professor of Computer Science Northwestern University
  Research Areas: Bayesian decision theory, Uncertainty Quantification, AI for Decision-Making
- Dr. Hamed Hassani, Associate Professor of Electrical and Systems Engineering at the University of Pennsylvania
  Research Areas: Trustworthy ML, Information Theory
- Dr. Amine Bennouna, Assistant Professor of Operations at the Kellogg School of Management, Northwestern University
  Research Areas: Data-driven Decision Making, Optimization
- Dr. Eric Zelikman, CEO and Co-Founder of Humans&
  Research Areas: Large Language Models, AI for Science
- Dr. Sin Yu Bonnie Ho, Senior Research Scientist at Microsoft Health and Life Sciences
  Research Areas: AI for Decision-Making, Patient Care

Attendance and Schedule. Attendance We anticipate an audience size of 150+. To cultivate a vibrant audience for our workshop, we are implementing several strategies: including leveraging social media platforms (Twitter/X, Bluesky), academic networks (school promotion), and industry partnerships to disseminate information and generate interest. We are committed to reaching underrepresented groups in machine learning and healthcare, ensuring invitations and promotional materials are circulated through affinity groups (WIML, Black in AI, and others). *Events* We will host two poster sessions and two rounds of spotlight talks for students to present their work. Moreover, as an icebreaker, we plan to play two rounds of interactive events with our atendees, which is refered as *Human/AI Game*. TODO: describe high level plan Tentative Schedule: 8:30–9:00 *Breakfast + Human/AI Game #1 + Opening Remarks* | 9:00–9:30 *Invited Talk #1* | 9:30–10:00 *Invited Talk #2* | 10:00–11:00 *Poster Session #1 + Coffee Break (discussion time)* | 11:00–11:30 *Invited Talk #3* | 11:30–12:00 *Invited Talk #4* | 12:00–12:30 *Spotlight Oral Presentations #1 (three total, 10 min each)* | 12:30–1:15 *Networking/Mentorship Lunch (discussion time)* | 1:15–1:30 *Human/AI Game #2* | 1:30–2:00 *Invited Talk #5* | 2:00–2:30 *Invited Talk #6* | 2:30–3:30 *Poster Session #2 + Coffee Break (discussion time)* | 3:30–4:00 *Spotlight Oral Presentations #2 (three total, 10 min each)* | 4:00–4:30 *Invited Talk #7* | 4:30–5:00 *Interactive Panel with Invited Speakers (all attendees; discussion time)* | 5:00–5:05 *Closing Remarks*

Sponsorships. We are actively seeking sponsorship from both academia, such as UPenn IDEAS, UPenn ASSET, JHU Data Science and AI Institute and industry, such as Google DeepMind, Microsoft, and Apple. The obtained funds will be used as awards to students who submitted to our workshop and achieved spotlight (Top 1-2% submission). With additional funds, we also seek to provide travel grants to students from underrepresented groups or students experiencing financial difficulty.

Workshop Submissions and Timelines. Paper Submission: We will invite submissions that explores the topic of HAI. We accept extended abstracts of up to 4 pages (excluding references). Accepted papers will be presented as in-person posters. There are no formal proceedings, but accepted papers will be publicly listed online unless authors opt out for non-archival presentation. We will also allow submissions from accepted conference/journal papers up to a year ago; such submissions

will not go under a review process but directly determined by Area Chairs for acceptance, determined based on relevance to the workshop. Review Process: We will use the OpenReview system to enforce the NeurIPS Code of Conduct as well as this following policy: any reviewer (including organizers) cannot assess any submission from someone (1) who has been a colleague with the reviewer within the same organization in the past 3 years; (2) who has co-authored publications with the reviewer in the past 3 years; (3) is currently from the same institution as the submitting authors. We will make sure to recruit reviewers from a wide variety of institutions and training levels to ensure the review process is unbiased. Lastly, only unpublished work will be accepted for workshop proceedings. Timeline: *Call for papers*: March 22 | *Submission deadline*: April 24 | *Reviewing period*: April 26 - May 10 |*Notification of acceptance*: May 15 | *Workshop Date*: July

Organizers

This is the lead organizing team, responsible for the majority of workshop planning and coordination. We have a variety of workshop-organizing experience, with organizers that are well-experienced to junior researchers establishing their academic networks. Designated Contact: Kwan Ho Ryan Chan (ryanckh@seas.upenn.edu) and Sima Noorani (nooranis@seas.upenn.edu)

Kwan Ho Ryan Chan is a fifth-year Electrical and Systems Engineering Ph.D. Candidate at University of Pennsylvania, advised by Dr. René Vidal. His work intersects the theory and application of human-AI collaboration, such as building algorithms for uncertainty quantification and decision-making with Large Language Models. He is a recepient of the NSF Graduate Research Fellowship, Penn Engineering Dean's Fellowship and UPenn AWS-ASSET Fellowship. Previously, he interned at Amazon AWS AI Labs (2026) and AI/ML Health AI team at Apple (2024). He received his Bachelors of Art in Applied Mathematics from University of California, Berkeley. In the past, he helped with organization of conferences including Conference on Lifelong Learning Agents 2025, DeepMath Conference 2024, as well as volunteered at International Conference on Learning Representations 2022.

Sima Noorani Suspendisse vitae elit. Aliquam arcu neque, ornare in, ullamcorper quis, commodo eu, libero. Fusce sagittis erat at erat tristique mollis. Maecenas sapien libero, molestie et, lobortis in, sodales eget, dui. Morbi ultrices rutrum lorem. Nam elementum ullamcorper leo. Morbi dui. Aliquam sagittis. Nunc placerat. Pellentesque tristique sodales est. Maecenas imperdiet lacinia velit. Cras non urna. Morbi eros pede, suscipit ac, varius vel, egestas non, eros. Praesent malesuada, diam id pretium elementum, eros sem dictum tortor, vel consectetuer odio sem sed wisi.

Chris Chiu Suspendisse vitae elit. Aliquam arcu neque, ornare in, ullamcorper quis, commodo eu, libero. Fusce sagittis erat at erat tristique mollis. Maecenas sapien libero, molestie et, lobortis in, sodales eget, dui. Morbi ultrices rutrum lorem. Nam elementum ullamcorper leo. Morbi dui. Aliquam sagittis. Nunc placerat. Pellentesque tristique sodales est. Maecenas imperdiet lacinia velit. Cras non urna. Morbi eros pede, suscipit ac, varius vel, egestas non, eros. Praesent malesuada, diam id pretium elementum, eros sem dictum tortor, vel consectetuer odio sem sed wisi.

Jacopo Teneggi Suspendisse vitae elit. Aliquam arcu neque, ornare in, ullamcorper quis, commodo eu, libero. Fusce sagittis erat at erat tristique mollis. Maecenas sapien libero, molestie et, lobortis in, sodales eget, dui. Morbi ultrices rutrum lorem. Nam elementum ullamcorper leo. Morbi dui. Aliquam sagittis. Nunc placerat. Pellentesque tristique sodales est. Maecenas imperdiet lacinia velit. Cras non urna. Morbi eros pede, suscipit ac, varius vel, egestas non, eros. Praesent malesuada, diam id pretium elementum, eros sem dictum tortor, vel consectetuer odio sem sed wisi.

Hyewon Jeong Suspendisse vitae elit. Aliquam arcu neque, ornare in, ullamcorper quis, commodo eu, libero. Fusce sagittis erat at erat tristique mollis. Maecenas sapien libero, molestie et, lobortis in, sodales eget, dui. Morbi ultrices rutrum lorem. Nam elementum ullamcorper leo. Morbi dui. Aliquam sagittis. Nunc placerat. Pellentesque tristique sodales est. Maecenas imperdiet lacinia velit. Cras non urna. Morbi eros pede, suscipit ac, varius vel, egestas non, eros. Praesent malesuada, diam id pretium elementum, eros sem dictum tortor, vel consectetuer odio sem sed wisi.

Dr. Aditya Chattobadhyay is a Research Scientists at Amazon AWS AI Labs, supervised by Dr. Stafano Soatto. *The content of this proposal does not relate to Aditya Chattobadhyay's position at Amazon.* TODO: please say you organized tutorial Suspendisse vitae elit. Aliquam arcu neque, ornare in, ullamcorper quis, commodo eu, libero. Fusce sagittis erat at erat tristique mollis. Maecenas sapien libero, molestie et, lobortis in, sodales eget, dui. Morbi ultrices rutrum lorem. Nam elementum ullamcorper leo. Morbi dui. Aliquam sagittis. Nunc placerat. Pellentesque tristique sodales est. Maecenas imperdiet lacinia velit. Cras non urna. Morbi eros pede, suscipit ac, varius vel, egestas non, eros. Praesent malesuada, diam id pretium elementum, eros sem dictum tortor, vel consectetuer odio sem sed wisi.

Advisory Board

Dr. René Vidal is the Penn Integrates Knowledge and Rachleff University Professor of Electrical and Systems Engineering & Radiology and the Director of the Center for Innovation in Data Engineering and Science (IDEAS) at the University of Pennsylvania. He is also an Amazon Scholar, an Affiliated Chief Scientist at NORCE, and a former Associate Editor in Chief of TPAMI. His current research focuses on the foundations of deep learning and trustworthy AI and its applications in computer vision and biomedical data science. His lab has made seminal contributions to motion segmentation, action recognition, subspace clustering, matrix factorization, deep learning theory, interpretable AI, and biomedical image analysis. He is an ACM Fellow, AIMBE Fellow, IEEE Fellow, IAPR Fellow and Sloan Fellow, and has received numerous awards for his work, including the IEEE Edward J. McCluskey Technical Achievement Award, D'Alembert Faculty Award, J.K.

Aggarwal Prize, ONR Young Investigator Award, NSF CAREER Award as well as best paper awards in machine learning, computer vision, signal processing, controls, and medical robotics.

Dr. George Pappas is the UPS Foundation Professor at the Department of Electrical and Systems Engineering at the University of Pennsylvania. He also holds a secondary appointment in the Departments of Computer and Information Sciences, as well as Mechanical Engineering and Applied Mechanics. He currently serves as the Associate Dean for Research and Innovation in the School of Engineering and Applied Science and as the Director of the Raj and Neera Singh program in Artificial Intelligence. Pappas's research focuses on control systems, robotics, autonomous systems, formal methods, and machine learning for safe and secure cyber-physical systems. He has received numerous awards, including the NSF PECASE, the Antonio Ruberti Young Researcher Prize, the George S. Axelby Award, the O. Hugo Schuck Best Paper Award, and the George H. Heilmeier Faculty Excellence Award. Pappas has mentored more than fifty students and postdocs, now faculty in leading universities worldwide. He is a Fellow of IEEE, IFAC, and was elected to the National Academy of Engineering in 2024.

Program Committee

This is the list of Program Committee members who have all agreed to review the workshop papers.

Yuyan Ge (University of Pennsylvania), Tianjiao Ding (University of Pennsylvania), Darshan Thaker (University of Pennsylvania), Beepul Bharti (Johns Hopkins University), Andrea Wynn (Johns Hopkins University), Jie Gao (Johns Hopkins University), Zheng Zhang (Notre Dame University), Drew Prinster (Johns Hopkins University), Natalie Collina (University of Pennsylvania).

References

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

[2] Chinmaya Andukuri, Jan-Philipp Fränken, Tobias Gerstenberg, and Noah D Goodman. Star-gate: Teaching language models to ask clarifying questions. *arXiv preprint arXiv:2403.19154*, 2024.

[3] Kwan Ho Ryan Chan, Yuyan Ge, Edgar Dobriban, Hamed Hassani, and René Vidal. Conformal information pursuit for interactively guiding large language models. *arXiv preprint arXiv:2507.03279*, 2025.

[4] Christopher Chiu, Silviu Pitis, and Mihaela van der Schaar. Simulating viva voce examinations to evaluate clinical reasoning in large language models. *arXiv preprint arXiv:2510.10278*, 2025.

[5] Natalie Collina, Surbhi Goel, Varun Gupta, and Aaron Roth. Tractable agreement protocols. In *Proceedings of the 57th Annual ACM Symposium on Theory of Computing*, pages 1532–1543, 2025.

[6] Natalie Collina, Surbhi Goel, Aaron Roth, Emily Ryu, and Mirah Shi. Emergent alignment via competition. *arXiv preprint arXiv:2509.15090*, 2025.

[7] Natalie Collina, Ira Globus-Harris, Surbhi Goel, Varun Gupta, Aaron Roth, and Mirah Shi. Collaborative prediction: Tractable information aggregation via agreement. In *Proceedings of the 2026 Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 4712–4798. SIAM, 2026.

[8] Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

[9] Ziyang Guo, Yifan Wu, Jason D Hartline, and Jessica Hullman. A decision theoretic framework for measuring ai reliance. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, pages 221–236, 2024.

[10] Kunal Handa, Yarin Gal, Ellie Pavlick, Noah Goodman, Jacob Andreas, Alex Tamkin, and Belinda Z Li. Bayesian preference elicitation with language models. *arXiv preprint arXiv:2403.05534*, 2024.

[11] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

[12] Xiong Junwu, Xiaoyun Feng, YunZhou Shi, James Zhang, Zhongzhou Zhao, and Wei Zhou. Digital human interactive recommendation decision-making based on reinforcement learning. *arXiv preprint arXiv:2210.10638*, 2022.

[13] Been Kim, John Hewitt, Neel Nanda, Noah Fiedel, and Oyvind Tafjord. Because we have llms, we can and should pursue agentic interpretability. *arXiv preprint arXiv:2506.12152*, 2025.

[14] Belinda Z Li, Alex Tamkin, Noah Goodman, and Jacob Andreas. Eliciting human preferences with language models. *arXiv preprint arXiv:2310.11589*, 2023.

[15] Stella Li, Vidhisha Balachandran, Shangbin Feng, Jonathan Ilgen, Emma Pierson, Pang Wei W Koh, and Yulia Tsvetkov. Mediq: Question-asking llms and a benchmark for reliable interactive clinical reasoning. *Advances in Neural Information Processing Systems*, 37:28858–28888, 2024.

[16] Jessy Lin, Nicholas Tomlin, Jacob Andreas, and Jason Eisner. Decision-oriented dialogue for human-ai collaboration. *Transactions of the Association for Computational Linguistics*, 12:892–911, 2024.

[17] Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*, 2024.

[18] Sima Noorani, Shayan Kiyani, George Pappas, and Hamed Hassani. Human-ai collaborative uncertainty quantification. *arXiv preprint arXiv:2510.23476*, 2025.

[19] 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.

[20] Wasu Top Piriyakulkij, Volodymyr Kuleshov, and Kevin Ellis. Active preference inference using language models and probabilistic reasoning. *arXiv preprint arXiv:2312.12009*, 2023.

[21] Allen Z Ren, Jaden Clark, Anushri Dixit, Masha Itkina, Anirudha Majumdar, and Dorsa Sadigh. Explore until confident: Efficient exploration for embodied question answering. *arXiv preprint arXiv:2403.15941*, 2024.

[22] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022.

[23] Erzhuo Shao, Yifang Wang, Yifan Qian, Zhenyu Pan, Han Liu, and Dashun Wang. Sciscigpt: advancing human–ai collaboration in the science of science. *Nature Computational Science*, pages 1–15, 2025.

[24] Kimi Team, Yifan Bai, Yiping Bao, Guanduo Chen, Jiahao Chen, Ningxin Chen, Ruijue Chen, Yanru Chen, Yuankun Chen, Yutian Chen, et al. Kimi k2: Open agentic intelligence. *arXiv preprint arXiv:2507.20534*, 2025.

[25] Wen-Fan Wang, Chien-Ting Lu, Nil Ponsa i Campanyà, Bing-Yu Chen, and Mike Y Chen. Aideation: Designing a human-ai collaborative ideation system for concept designers. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, pages 1–28, 2025.

[26] An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

[27] Ling Yang, Zhilong Zhang, Yang Song, Shenda Hong, Runsheng Xu, Yue Zhao, Wentao Zhang, Bin Cui, and Ming-Hsuan Yang. Diffusion models: A comprehensive survey of methods and applications. *ACM computing surveys*, 56(4): 1–39, 2023.

[28] Sherry Yang, KwangHwan Cho, Amil Merchant, Pieter Abbeel, Dale Schuurmans, Igor Mordatch, and Ekin Dogus Cubuk. Scalable diffusion for materials generation. *arXiv preprint arXiv:2311.09235*, 2023.