CS59200 Human-AI Interaction

Instructor: Ming Yin, Assistant Professor of Computer Science

Email: mingyin@purdue.edu

Lecture: 15:30-18:20 M @ LWSN 1106

Instructional Modality: Face-to-Face

Course Credits: 3.0

Prerequisites: Basic knowledge of human-computer interaction and artificial intelligence is recommended.

Course Description

Human-AI interaction sits at the intersection of human-computer interaction and artificial intelligence, and relate to psychology, communication, cognitive science, and design. In this course, we will focus on learning how to develop empirical understandings of humans' interactions with AI systems, and how to incorporate user-centered design principles to design AI systems that can enable effective interactions between people and the systems. This course starts with a brief review of fundamentals of human cognition and artificial intelligence, as well as a discussion of user-centered design lifecycle and general principles for designing human-AI interactions. Then, we will delve into a wide range of specific topics on human-AI interaction, including how to design explainable, trustworthy, fair and ethical AI systems, how to enable effective human-AI collaboration and teaming, and what new opportunities and challenges do the rise of large language models bring to human-AI interaction.

Grading

- Assignment: 10%
- Reading responses: 15%
- Class participation and discussion: 10%
- Paper presentation: 20%
- Final project: 45%

Course Schedule

Week 1 (Jan 8): Introduction to Human-AI Interaction and Course Overview

Week 2 (Jan 15): *No class (Martin Luther King Day)*

Week 3 (Jan 22): Fundamentals of Human Cognition and Artificial Intelligence

Optional readings:

- Evans, Heuristic and Analytic Processes in Reasoning (British Journal of Psychology 1984)
- Kahneman and Tversky, Judgment under Uncertainty: Heuristics and Biases (Science 1974)
- Chapter 3 - Mental Models and User Models from Handbook of Human-Computer Interaction
- Russell and Norvig. Artificial Intelligence: A Modern Approach, 4th edition, 2020

Week 4 (Jan 29): Human-Centered Design

Optional readings:

- Sharp, Preece, and Rogers. Interaction Design: Beyond Human-Computer Interaction, 2015
- Norman. The Design of Everyday Things: Revised and Expanded Edition, 2013
- Martin. Doing Psychology Experiments, 2007

Week 5 (Feb 5): Design Principles and Guidelines for Human-AI Interaction

Required readings:

- Amershi et al., Guidelines for Human-AI Interaction (CHI 2019)

Optional readings:

- Horvitz, Principles of Mixed-Initiative User Interfaces (CHI 1999)
- Shneiderman. Human-Centered AI, Oxford University Press, 2022
- Russell. Human Compatible: Artificial Intelligence and the Problem of Control, 2019

Week 6 (Feb 12): Explainable AI Part I (Definitions, Methods, and Human-Centered Evaluations)

Required readings:

- Ribeiro et al., "Why Should I Trust You?": Explaining the Predictions of Any Classifier (KDD 2016)
- Wang and Yin, Effects of Explanations in AI-Assisted Decision Making: Principles and Comparisons (ACM Transactions on Interactive Intelligent Systems, 2022)
- Cheng et al., Explaining Decision-Making Algorithms through UI: Strategies to Help Non-Expert Stakeholders (CHI, 2019)

Optional readings:

- Lakkaraju et al., Interpretable Decision Sets: A Joint Framework for Description and Prediction (KDD 2016)
- Miller, Explanation in artificial intelligence: Insights from the social sciences (Artificial Intelligence 2019)
- Guidotti et al., A Survey of Methods for Explaining Black Box Models (ACM Computing Surveys, August 2018)
- Molnar, Interpretable Machine Learning: A Guide for Making Black Box Models Explainable, 2020.
- Liao et al., Questioning the AI: Informing Design Practices for Explainable AI User Experiences (CHI 2020)
- Yang et al., How Do Visual Explanations Foster End Users' Appropriate Trust in Machine Learning? (IUI 2020)
- Bucinca et al., Proxy tasks and subjective measures can be misleading in evaluating explainable AI systems (IUI 2020)

- Poursabzi-Sangdeh et al., Manipulating and Measuring Model Interpretability (CHI 2021)
- Bansal et al., Does the Whole Exceed its Parts? The Effect of AI Explanations on Complementary Team Performance (CHI 2021)
- Vasconcelos et al., Explanations Can Reduce Overreliance on AI Systems During Decision-Making (CSCW 2023)
- Chen et al., Machine Explanations and Human Understanding (FAccT 2023)

Week 7 (Feb 19): Explainable AI Part II (Intervention Designs)

Required readings:

- Abdul et al., COGAM: Measuring and Moderating Cognitive Load in Machine Learning Model Explanations (CHI 2020)
- Lai et al., Selective Explanations: Leveraging Human Input to Align Explainable AI (CSCW 2023)

Optional readings:

- Wang et al., Designing Theory-Driven User-Centric Explainable AI (CHI 2019)
- Ehsan et al., Expanding Explainability: Towards Social Transparency in AI systems (CHI 2021)
- Fel et al., Harmonizing the Object Recognition Strategies of Deep Neural Networks with Humans (NeurIPS 2022)
- Nguyen et al., Visual Correspondence-based Explanations Improve AI Robustness and Human-AI Team Accuracy (NeurIPS 2022)
- Zhang and Lim, Towards Relatable Explainable AI with the Perceptual Process (CHI 2022)
- Gajos and Mamykina, Do People Engage Cognitively with AI? Impact of AI Assistance on Incidental Learning (IUI 2022)
- Chong et al., Human confidence in artificial intelligence and in themselves: The evolution and impact of confidence on adoption of AI advice (Computers in Human Behavior, 2022)
- Danry et al., Don't Just Tell Me, Ask Me: AI Systems that Intelligently Frame Explanations as Questions Improve Human Logical Discernment Accuracy over Causal AI explanations (CHI 2023)
- Slack et al., Fooling LIME and SHAP: Adversarial Attacks on Post hoc Explanation Methods (AIES 2020)
- Slack et al., Explaining Machine Learning Models with Interactive Natural Language Conversations Using TalkToModel (Nature Machine Intelligence, 2023)
- Miller, Explainable AI is Dead, Long Live Explainable AI!: Hypothesis-driven Decision Support using Evaluative AI (FAccT 2023)

Week 8 (Feb 26): Trust and Reliance on AI Part I (Empirical Studies and Computational Models)

Required readings:

- Rechkemmer and Yin. When Confidence Meets Accuracy: Exploring the Effects of Multiple Performance Indicators on Trust in Machine Learning Models (CHI 2022)
- Tejeda et al., AI-Assisted Decision-making: A Cognitive Modeling Approach to Infer Latent Reliance Strategies (Computational Brain & Behavior 2022)

Optional readings:

- Bansal et al., Beyond Accuracy: The Role of Mental Models in Human-AI Team Performance (HCOMP 2019)

- Bansal et al., Updates in Human-AI Teams: Understanding and Addressing the Performance/Compatibility Tradeoff (AAAI 2019)
- Zhang et al., Effect of confidence and explanation on accuracy and trust calibration in AI-assisted decision making (FAT* 2020)
- Nourani et al., The Role of Domain Expertise in User Trust and the Impact of First Impressions with Intelligent Systems (HCOMP 2020)
- Lu and Yin. Human Reliance on Machine Learning Models When Performance Feedback is Limited: Heuristics and Risks (CHI 2021)
- Guo and Yang, Modeling and Predicting Trust Dynamics in Human–Robot Teaming: A Bayesian Inference Approach (International Journal of Social Robotics, 2021)
- Azevedo-Sa et al., Real-Time Estimation of Drivers' Trust in Automated Driving Systems (International Journal of Social Robotics, 2021)
- Wang et al., Will You Accept the AI Recommendation? Predicting Human Behavior in AI-Assisted Decision Making (WWW 2022)
- Papenmeier et al., It's Complicated: The Relationship between User Trust, Model Accuracy and Explanations in AI (ACM Transactions on Computer-Human Interaction, 2022)
- Li et al., Modeling Human Trust and Reliance in AI-Assisted Decision Making: A Markovian Approach (AAAI 2023)
- Chen et al., Understanding the Role of Human Intuition on Reliance in Human-AI Decision-Making with Explanations (CSCW 2023)

Week 9 (Mar 4): Trust and Reliance on AI Part II (Intervention Designs)

Required readings:

- Ma et al., Who Should I Trust: AI or Myself? Leveraging Human and AI Correctness Likelihood to Promote Appropriate Trust in AI-Assisted Decision-Making (CHI 2023)
- Bansal et al., Is the Most Accurate AI the Best Teammate? Optimizing AI for Teamwork (AAAI 2021)

Optional readings:

- Bucinca et al., To Trust or to Think: Cognitive Forcing Functions Can Reduce Overreliance on AI in AI-assisted Decision-making (CSCW 2022)
- Rastogi et al., Deciding Fast and Slow: The Role of Cognitive Biases in AI-assisted Decision-making (CSCW 2022)
- Vodrahalli et al., Uncalibrated Models Can Improve Human-AI Collaboration (NeurIPS 2022)
- Benz and Rodriguez, Human-Aligned Calibration for AI-Assisted Decision Making (NeurIPS 2023)
- Cabrera et al., Improving Human-AI Collaboration With Descriptions of AI Behavior (CSCW 2023)
- Noti and Chen, Learning When to Advise Human Decision Makers (IJCAI 2023)
- Li et al., Strategic Adversarial Attacks in AI-assisted Decision Making to Reduce Human Trust and Reliance (IJCAI 2023)
- Inkpen et al., Advancing Human-AI Complementarity: The Impact of User Expertise and Algorithmic Tuning on Joint Decision Making (ACM Transactions on Computer-Human Interaction, 2023)

Week 10 (Mar 11): *No class (Spring break)*

Week 11 (Mar 18): Bias and Fairness in AI Part I (Definitions and Methods)

Required reading:

- Angwin et al., Machine Bias. 2016
- Srinivasan and Chander, Biases in AI Systems (Communications of ACM, 2021)
- Zafar et al., Fairness Beyond Disparate Treatment & Disparate Impact: Learning Classification without Disparate Mistreatment (WWW 2017)

Optional reading:

- Hardt et al., Equality of opportunity in supervised learning (NeurIPS 2016)
- Caliskan et al., Semantics Derived Automatically from Language Corpora Contain Human-Like Biases (Science 2017)
- Kusner et al., Counterfactual Fairness (NeurIPS 2017)
- Otterbacher et al., Competent Men and Warm Women: Gender Stereotypes and Backlash in Image Search Results (CHI 2017)
- Hube et al., Understanding and Mitigating Worker Biases in the Crowdsourced Collection of Subjective Judgments (CHI 2019)
- Bellamy et al., AI Fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias (IBM Journal of Research and Development, 2019)
- Karimi et al., Algorithmic Recourse: from Counterfactual Explanations to Interventions (FaccT 2021)
- Mehrabi et al., A Survey on Bias and Fairness in Machine Learning (ACM Computing Surveys 2021)
- Mitchell et al., Algorithmic Fairness: Choices, Assumptions, and Definitions (Annual Review of Statistics and Its Application, 2021)
- Sap et al., Annotators with attitudes: How annotator beliefs and identities bias toxic language detection (NAACL 2022)

Week 12 (Mar 25): Bias and Fairness in AI Part II (Empirical Studies and Intervention Designs)

Required readings:

- Wang et al., Factors Influencing Perceived Fairness in Algorithmic Decision-Making: Algorithm Outcomes, Development Procedures, and Individual Differences (CHI 2019)
- Green and Chen. Disparate Interactions: An Algorithm-in-the-Loop Analysis of Fairness in Risk Assessments (FAT* 2019)

Optional readings:

- Liu et al., Delayed impact of fair machine learning (ICML 2018)
- Grgic-Hlaca et al., Human Perceptions of Fairness in Algorithmic Decision Making: A Case Study of Criminal Risk Prediction (WWW 2018)
- Srivastava et al., Mathematical Notions vs. Human Perception of Fairness: A Descriptive Approach to Fairness for Machine Learning (KDD 2019)
- Saxena et al., How Do Fairness Definitions Fare?: Examining Public Attitudes Towards Algorithmic Definitions of Fairness (AIES 2019)
- Dodge et al., Explaining models: an empirical study of how explanations impact fairness judgment (IUI 2019)
- Zhang et al., Group Retention when Using Machine Learning in Sequential Decision Making: the Interplay between User Dynamics and Fairness (NeurIPS 2019)
- Zhang et al., How do fair decisions fare in long-term qualification? (NeurIPS 2020)

- Gemalmaz and Yin, Accounting for Confirmation Bias in Crowdsourced Label Aggregation (IJCAI 2021)
- Cheng et al., How Child Welfare Workers Reduce Racial Disparities in Algorithmic Decisions (CHI 2022)
- Gordon et al., Jury learning: Integrating dissenting voices into machine learning models (CHI 2022)
- Wang and Yin, The Effects of AI Biases and Explanations on Human Decision Fairness: A Case Study of Bidding in Rental Housing Markets (IJCAI 2023)

Week 13 (Apr 1): Human-AI Collaboration and Teaming

Required readings:

- Lai et al., Human-AI Collaboration via Conditional Delegation: A Case Study of Content Moderation (CHI 2022)
- Hong et al., Learning to influence human behavior with offline reinforcement learning (NeurIPS 2023)

Optional readings:

- Madras et al., Predict Responsibly: Improving Fairness and Accuracy by Learning to Defer (NeurIPS 2018)
- Nushi et al., Towards Accountable AI: Hybrid Human-Machine Analyses for Characterizing System Failure (HCOMP 2018)
- Carroll et al., On the utility of learning about humans for human-AI coordination (NeurIPS 2019)
- Chakraborti et al., Balancing Explicability and Explanations for Human-Aware Planning (IJCAI 2019)
- Gennatas et al., Expert-augmented Machine Learning (PNAS 2020)
- Xiao et al., Fresh: Interactive reward shaping in high-dimensional state spaces using human feedback (AAMAS 2020)
- Wilder et al., Learning to Complement Humans (IJCAI 2020)
- Gao et al., Human-AI Collaboration with Bandit Feedback (IJCAI 2021)
- Steyvers et al., Bayesian modeling of human–AI complementarity (PNAS 2022)
- Callaway et al., Leveraging artificial intelligence to improve people's planning strategies (PNAS 2022)
- Schelble et al., Let's Think Together! Assessing Shared Mental Models, Performance, and Trust in Human-Agent Teams (GROUP 2022)

Week 14 (Apr 8): Human Interaction with Large Language Models

Required readings:

- Zamfiresce-Pereira et al., Why Johnny Can't Prompt: How Non-AI Experts Try (and Fail) to Design LLM Prompts (CHI 2023)
- Jakesch et al., Co-Writing with Opinionated Language Models Affects Users' Views (CHI 2023)

Optional readings:

- Wu et al., AI Chains: Transparent and Controllable Human-AI Interaction by Chaining Large Language Model Prompts (CHI 2022)
- Yuan et al., Wordcraft: Story Writing With Large Language Models (IUI 2022)

- Noy and Zhang, Experimental evidence on the productivity effects of generative artificial intelligence (Science 2023)
- Argyle et al., Leveraging AI for democratic discourse: Chat interventions can improve online political conversations at scale (PNAS 2023)
- Wang et al., PopBlends: Strategies for Conceptual Blending with Large Language Models (CHI 2023)
- Chung et al., Increasing Diversity While Maintaining Accuracy: Text Data Generation with Large Language Models and Human Interventions (ACL 2023)
- Fok et al., Scim: Intelligent Skimming Support for Scientific Papers (IUI 2023)
- Rastogi et al., Supporting Human-AI Collaboration in Auditing LLMs with LLMs (AIES 2023)
- Xiao et al., Supporting Qualitative Analysis with Large Language Models: Combining Codebook with GPT-3 for Deductive Coding (IUI Companion 2023)

Week 15 (Apr 15): AI, Ethics, and Society

Required readings:

- Awad et al., The Moral Machine Experiment (Nature 2018)
- Gabriel, Artificial Intelligence, Values, and Alignment (Minds and Machines 2020)

Optional readings:

- Conitzer et al., Moral Decision Making Frameworks for Artificial Intelligence (AAAI 2017)
- Zhu et al., Value-Sensitive Algorithm Design: Method, Case Study, and Lessons (CSCW 2018)
- Smith et al., Keeping Community in the Loop: Understanding Wikipedia Stakeholder Values for Machine Learning-Based Systems (CHI 2020)
- Tolmeijer et al., Capable but Amoral? Comparing AI and Human Expert Collaboration in Ethical Decision Making (CHI 2022)
- Zhang et al., Artificial intelligence and moral dilemmas: Perception of ethical decision-making in AI (Journal of Experimental Social Psychology, 2022)
- Narayanan et al., How does Value Similarity affect Human Reliance in AI-Assisted Ethical Decision Making? (AIES 2023)
- Rezwana and Maher, User Perspectives on Ethical Challenges in Human-AI Co-Creativity: A Design Fiction Study (C&C 2023)

Week 16 (Apr 22): Final project presentation

Paper Reading, Presentation and Discussion

Many classes in this course consist of paper reading, presentation and discussion. Specifically, in a typical class, we will cover 2-3 papers on one topic, and each paper will be assigned to one student so that the student will present the paper and lead the discussion. Responsibility of presenters of one class include:

- Read the assigned paper, and at least one optional paper, for that class.
- After discussing with the instructor (one week before the class), post 2-3 conversation-provoking questions related to the required paper(s) of that class.
- Give a presentation in class, which should review all required paper(s) of that class, and also briefly introduce the optional paper(s) that they have read.
- Lead the discussion in class.

Responsibility of non-presenters of one class include:

- Read all required paper(s) for that class.
- Before class, provide reading responses to all questions that the presenters of that class post.
- Participate in the discussion in class.

Final Project

Final project serves as an opportunity for students to get hands-on experience in human-centered computing research. Projects are open-ended; sample projects include:

- Design and conduct online experiments to investigate how various factors of AI systems affect trust and adoption of the systems
- Construct new interpretable ML / fair ML methods (e.g., for innovative use cases / types of data) and compare its effectiveness with existing methods through user studies.
- Design new interactive AI systems for specific applications
- Explore how large language models can be used to different contexts to augment humans' abilities and creativity, and how humans can help improve the outputs of large language models
- Examine how to model humans' behavior in their interactions with AI, and how to adjust the designs of AI systems to accommodate these human behaviors

Students are also encouraged to connect the final project with their own research.

Students can complete the project either individually or in a group of two. Tasks related to the final project include:

- Submit a project proposal which identifies the problem that the project aims to solve.
- Submit a mid-term report on the project progress and get feedback.
- Give a final presentation on the project in class, reporting the results of the project.
- Submit a final project report summarizing the project.

More detailed instruction on the final project will be provided through project guidelines.