CS592 Human-AI Interaction

Instructor: Tianyi Zhang, Assistant Professor of Computer Science

Email: To be filled

Office: To be filled

Lecture: To be filled

Office Hours: Tue and Thurs 2pm-3pm (tentative)

Course Description

Have you ever wondered about these:

- What is the role of humans in the future of AI?
- Will programming jobs no longer exist because of large language models like GPT-3?[1]
- How far are we from the "black art" of natural language programming as Dijkstra called it 40 years ago?[2]
- Why does IBM suddenly seek to sell Watson Health, their AI for healthcare division, after 10 years of huge investment?[3]
- Self-driving cars are coming, but are we ready?
- How can humans efficiently give feedback to AI and correctify its mistakes?
- How will humans and AI evolve together in the next decade?

This course will help you answer those questions. You will read and discuss research papers in human-AI interaction, including but not limited to research topics about (1) AI-based systems and applications working with---or clashing against---the strengths and weaknesses of human cognition, (2) how to design interactive, human-in-the-loop approaches that achieve human-AI symbiosis, and (3) how to support interpretability, transparence, trust, and fairness in AI-based systems. Specifically, we will look into recent research advances in several trending domains such as "AI for code", healthcare, and autonomous driving.

Activities will include a small number of lectures, presentations of research papers, and discussion of relevant literature in each field. You should expect to present one or two research papers during the semester. You also need to write a one-paragraph paper

[1] OpenAI's GPT-3 Can Now Generate The Code For You. *Analytics India Magazine*, July 20, 2020.

[2] E. W. Dijkstra. On the foolishness of "natural language programming". In *Program Construction*, pages 51–53. Springer, 1979.

[3] L. Cooper and C. Lombardo. IBM Explores Sale of IBM Watson Health. *The Wall Street Journal*, Feb. 18, 2021.

review in the form of comments and questions and post it on Piazza before each paper discussion. There will be a course project, in which you will work in groups to design and carry out research projects related to human-AI interaction. Depending on the schedule, we will have one or two guest speakers to present their current research in human-AI interaction.

This course is designed to introduce research topics in human-AI interaction. You do not need to have a strong ML or HCI background to take this class. PhD students in AI/ML, HCI, SE/PL, NLP, Vision, Robotics, Security, and Visualization who are interested in human-AI interaction are encouraged to take this class. Masters students and advanced undergraduates, particularly those who wish to do research or write a thesis, are also welcome. If you are not sure about your qualification for this course, feel free to research out to the instructor via email.

Course Objectives

At the end of this course, students should be able to:

- identify and understand the problem statement, research questions, methods, findings, and contributions in a research paper
- critically assess the contributions of a paper
- design and implement interactive systems with AI components
- evaluate an interactive AI-based system, especially through user studies in the lab or on a crowdsourcing platform like Amazon Mechanical Turks
- know the style of academic writing, especially in HCI and Software Engineering
- make and deliver academic presentations to the public

Course Schedule

Week 1. Introduction to Human-AI Interaction

Lecture 1. Course description and the design argument framework

Lecture 2. An overview of human-AI interaction

- Amershi et al., Power to the People: The Role of Humans in Interactive Machine Learning (AI Magazine 2014)
- Dudley and Kristensson, A Review of User Interface Design for Interactive Machine Learning (TIIS 2018)

Paper presentation sign-up due this week

Week 2. Human Needs, Perceptions, and Experiences of Using AI (Part I)

Lecture 1. The needs, perceptions, and experiences of end-users

- Eiband et al., When People and Algorithms Meet: User-reported Problems in Intelligent Everyday Applications (IUI 2019)
- Luger and Sellen, “Like Having a Really Bad PA”: The Gulf between User Expectation and Experience of Conversational Agents (CHI 2016).

Optional reading:

- Frison et al., Why Do You Like To Drive Automated? (IUI 2019)
- Tullio et al., How it works: a field study of non-technical users interacting with an intelligent system (CHI 2007)
- Rader and Gray, Understanding User Beliefs About Algorithmic Curation in the Facebook News Feed (CHI 2015)
- Q. Vera Liao et al., All Work and No Play? Conversations with a Question-and-Answer Chatbot in the Wild (CHI 2018)

Lecture 2. The needs, perceptions, and experiences of software developers

- Weisz et al., Perfection Not Required? Human-AI Partnerships in Code Translation (IUI 2021)
- Xu et al., In-IDE Code Generation from Natural Language: Promise and Challenges (arXiv 2021)

Optional readings:

- Hellendoorn et al., When Code Completion Fails: a Case Study on Real-World Completions (ICSE 2019)
- Tao et al., Automatically generated patches as debugging aids: a human study (FSE 2014)
- Cambronero et al., Characterizing Developer Use of Automatically Generated Patches (VH/HCC 2019)

Week 3. Human Needs, Perceptions, and Experiences of Using AI (Part II)

Lecture 1. The needs, perceptions, and experiences of data scientists

- Hohman et al., Gamut: A Design Probe to Understand How Data Scientists Understand Machine Learning Models (CHI 2019)
- Kaur et al., Interpreting Interpretability: Understanding Data Scientists’ Use of Interpretability Tools for Machine Learning (CHI 2020)

Lecture 2. The needs, perceptions, and experiences of other domain experts

- Cai et al., "Hello AI": Uncovering the Onboarding Needs of Medical Practitioners for Human-AI Collaborative Decision-Making (CSCW 2019)
- Levy et al., Assessing the Impact of Automated Suggestions on Decision Making: Domain Experts Mediate Model Errors but Take Less Initiative (CHI 2021)

Optional readings:

- Khairat et al., Reasons For Physicians Not Adopting Clinical Decision Support Systems: Critical Analysis (JMIR 2018)
- Jacobs et al., Designing AI for Trust and Collaboration in Time-Constrained Medical Decisions: A Sociotechnical Lens (CHI 2021)

Week 4. Heuristics, Biases, and Mental Models of AI Agents

Lecture 1. Heuristics and biases in human decision making

- Kahneman and Tversky, Judgment under Uncertainty: Heuristics and Biases (Science 1974)
- Lu and Yin, Human Reliance on Machine Learning Models When Performance Feedback is Limited: Heuristics and Risks (CHI 2021)

Lecture 2. How will users' mental models impact their interaction with AI agents?

- Gero et al., Mental Models of AI Agents in a Cooperative Game Setting (CHI 2019)
- Bansal et al., Beyond Accuracy: On the Role of Mental Models in Human-AI Teams (HCOMP 2019)

Optional reading:

- Kocielnik, R., Amershi, S., and Bennett, P. Will You Accept an Imperfect AI? Exploring Designs for Adjusting End-User Expectations of AI Systems. (CHI 2019)

Week 5. Design Principles and Guidelines for Human-AI interaction

Lecture 1. Historical Perspectives of Human-AI Interaction Design

- Schneiderman and Maes, Direct Manipulation vs. Interface Agents (Interactions 1997)
- Horvitz, Principles of Mixed-Initiative Interaction (CHI 1999)

Optional reading:

- Licklider, Man-Computer Symbiosis (IRE Transactions on Human Factors in Electronics, 1960)

Lecture 2. A Contemporary View of Human-AI Interaction Design

1. Amershi et al., Guidelines for Human-AI Interaction (CHI 2019)
2. Heer, Agency plus automation: Designing artificial intelligence into interactive systems (PNAS 2019)

Optional readings:

- Yang et al., Re-examining Whether, Why, and How Human-AI Interaction Is Uniquely Difficult to Design (CHI 2020)

Quiz 1
Project proposal due this week

Week 6. Concrete Human-AI Interaction Designs (Part I)

Lecture 1. Conveying model confidence and uncertainty

- Verame et al., The effect of displaying system confidence information on the usage of autonomous systems for non-specialist applications: A lab study. (CHI 2016)
- Kay et al., When (ish) is My Bus? User-centered Visualizations of Uncertainty in Everyday, Mobile Predictive Systems (CHI 2016)

Optional reading:

- Antifakos et al., Towards improving trust in context-aware systems by displaying system confidence (MobileHCI 2005)

Lecture 2. Supporting model customization, refinement, and correction

- Li et al., Multi-Modal Repairs of Conversational Breakdowns in Task-Oriented Dialogs (UIST 2020)
- Koh et al., Concept Bottleneck Models (ICML 2020)

Optional reading:

- Kulesza et al., Principles of Explanatory Debugging to Personalize Interactive Machine Learning (IUI 2015)
- Cai et al. Human-Centered Tools for Coping with Imperfect Algorithms during Medical Decision-Making (CHI 2019)

Week 7. Concrete Human-AI Interaction Designs (Part II)

Lecture 1. Providing explanation and help users understand model behavior

- Cai et al., The effects of example-based explanations in a machine learning interface (IUI 2019)
- Schneider et al., ExplAIn Yourself! Transparency for Positive UX in Autonomous Driving (CHI 2021)

Optional reading:

- What AI can do for me: Evaluating Machine Learning Interpretations in Cooperative Play (IUI 2019)

Lecture 2. Actively eliciting and incorporating user feedback

- Siveraman et al., Active Inductive Logic Programming for Code Search (ICSE 2019)
- Yao et al., Interactive Semantic Parsing for If-Then Recipes via Hierarchical Reinforcement Learning **(**AAAI 2019)

Optional readings:

- Fogarty et al., CueFlik: interactive concept learning in image search (CHI 2008)

- Amershi et al., Overview based example selection in end user interactive concept learning (UIST 2009)
- Yao et al., Model-based Interactive Semantic Parsing: A Unified Framework and A Text-to-SQL Case Study (EMNLP 2019)

Quiz 2

Week 8. Augment AI to Cope with Limitations of Human Users

Lecture 1. Deal with limited attention and overreliance on AI

- Bucinca et al., To Trust or to Think: Cognitive Forcing Functions Can Reduce Overreliance on AI in AI-assisted Decision-making (CSCW 2021)
- Kulesza et al., Too Much, Too Little, or Just Right? Ways Explanations Impact End Users' Mental Models (VL/HCC 2013)

Optional reading:

- Croskerry, Cognitive Forcing Strategies in Clinical Decisionmaking (Annals of emergency medicine 2003)

Lecture 2. Resolve ambiguity in human intent and communication to AI

- Zhang et al., Interactive Program Synthesis by Augmented Examples (UIST 2020)
- Narita et al., Data-centric disambiguation for data transformation with programming-by-example (IUI 2021)

Optional readings:

- Mayer et al., User Interaction Models for Disambiguation in Programming by Example (UIST 2015)
- Pu et al., Program Synthesis with Pragmatic Communication (NeurIPS 2020)

Week 9. Interpretability and Explainability

Lecture 1. Example-based explanations and counterfactuals

- Wexler et al., The What-If Tool: Interactive Probing of Machine Learning Models (VAST 2019)
- Kim et al., Examples are not Enough, Learn to Criticize! Criticism for Interpretability (NIPS 2016)

Other readings:

- Byrne, Counterfactuals in Explainable Artificial Intelligence (XAI): Evidence from Human Reasoning (IJCAI 2019)

Lecture 2. Model-agnostic explanation and feature attribution

- Ribeiro et al., "Why Should I Trust You?": Explaining the Predictions of Any Classifier (KDD 2016)

- Olah et al., The Building Blocks of Interpretability (Distill 2018)

Optional readings:

- Molnar, Interpretable Machine Learning: A Guide for Making Black Box Models Explainable (2021)
- Liao et al., Questioning the AI: Informing Design Practices for Explainable AI User Experiences (CHI 2020)
- Bhatt et al., Explainable Machine Learning in Deployment (FAT 2020)

Quiz 3
Mid-point project summary due this week

Week 10. Interactive Visual Analytics for Machine Learning

Lecture 1. Visualization for understanding model behavior

- Strobelt et al., LSTMVis: A Tool for Visual Analysis of Hidden State Dynamics in Recurrent Neural Networks (TVCG 2017)
- Kahng et al. ACTIVIS: Visual Exploration of Industry-Scale Deep Neural Network Models (VAST 2017)

Optional reading:

- Strobelt et al., Seq2seq-Vis: A Visual Debugging Tool for Sequence-to-Sequence Models (TVCG 2018)
- Park et al., SANVis: Visual Analytics for Understanding Self-Attention Networks (VAST 2019)

Lecture 2. Visualization for model comparison and selection

- Xu et al., mTSeer: Interactive Visual Exploration of Models on Multivariate Time-series Forecast (CHI 2021)
- Yan et al., Visualizing Examples of Deep Neural Networks at Scale (CHI 2021)

Optional readings:

- Hohman et al., Visual Analytics in Deep Learning: An Interrogative Survey for the Next Frontiers (TVCG 2019)

Week 11. Reliability and Trust

Lecture 1. Principles and Human Perceptions

- Shneiderman, Human-Centered Artificial Intelligence: Reliable, Safe & Trustworthy (IJHCI 2020)
- Skirpan et al., What's at Stake: Characterizing Risk Perceptions of Emerging Technologies. (CHI 2018)

Optional readings:

- Dzindolet et al., The role of trust in automation reliance (International Journal of Human-Computer Studies 2003)
- Nicholas Diakopoulos, Algorithmic Accountability (Digital Journalism 2015)

Lecture 2. Trust Calibration

- Zhang et al., Effect of confidence and explanation on accuracy and trust calibration in AI-assisted decision making (FAT 2020)
- Häuslschmid et al., Supporting Trust in Autonomous Driving (IUI 2017)

Optional reading:

- Okamura and Yamada, Adaptive trust calibration for human-AI collaboration (PLOS ONE)

Quiz 4

Week 12. AI Ethics, Fairness, and Equity

Lecture 1. AI bias

- Angwin et al., Machine bias: There's software used across the country to predict future criminals, and it's biased against blacks (ProPublica 2016)
- Buolamwini and Gebru, Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. FACCT 2018

Other readings:

- Verma and Rudin, Fairness definitions explained (FairWare 2018)
- Mehrabi et al., A survey on bias and fairness in machine learning (arXiv 2019)
- Caliskan et al., Semantics derived automatically from language corpora contain human-like biases (Science 2017)

Lecture 2. Bias Detection and Fairness Testing

- Galhotra et al., Fairness testing: testing software for discrimination (ESEC/FSE 2017)
- Cabrera et al., FairVis: Visual Analytics for Discovering Intersectional Bias in Machine Learning (VAST 2019)

Optional reading:

- Holstein et al., Improving Fairness in Machine Learning Systems: What Do Industry Practitioners Need (CHI 2019)

Week 13. Human-AI Co-creation in Different Domains

Lecture 1

- (Writing) Gehrmann et al., Visual Interaction with Deep Learning Models through Collaborative Semantic Inference (VAST 2019)

- (UI Design) Swearngin et al., Scout: Rapid Exploration of Interface Layout Alternatives through High-Level Design Constraints (CHI 2020)

Optional reading:

- (Writing) Clark et al., Creative writing with a machine in the loop: Case studies on slogans and stories (IUI 2018)

Lecture 2

- (Fact checking) Nguyen et al., Believe it or not: Designing a Human-AI Partnership for Mixed-Initiative Fact-Checking (UIST 2018)
- (Music composition) Louie et al., Novice-AI Music Co-Creation via AI-Steering Tools for Deep Generative Models (CHI 2020)

Optional reading:

- (Video Games) Guzdial et al., Friend, Collaborator, Student, Manager: How Design of an AI-Driven Game Level Editor Affects Creators (CHI 2019)

Quiz 5

Week 14. Thanksgiving Week (No Class)

Week 15. Final Project Presentations

Final project report due this week

Paper Reading Assignment, Presentation, and Discussion

For each lecture, you should expect to read two research papers on a specific topic in Human-AI Interaction. I will provide some optional readings related to the topic but you are not required to read them. The optional readings are mostly for students who are particularly interested in the topic or who are doing a course project in the topic.

For each required paper reading, you need to submit a short paper review (one or two paragraphs) in the form of questions and comments on Piazza before the class. The grading of your paper review will depend on the overall quantity and quality of your questions and comments. As you read a paper or write your review, focus on the following perspectives.

- **Motivation of the work.** If the paper presents a new tool, who are the target users? Do they really need such a tool? What pain points does this tool address for those users? If the paper presents an empirical study, what are the research questions this study aim to answer? How important are these studies? Who will care about the findings and why should they care?

- **Novelty and significance of the work.** What is new here? What are the main contributions of the paper? What did you find most interesting?
- **Limitations, flaws, and blind spots.** Are there any unrealistic or false assumptions about the target users or the approach? Are there flaws or mistakes in the tool design, technical approach, or the study design?
- **Future work.** How would you improve on this work? Does this paper inspire any new ideas in your own research?

Depending on the number of students enrolled in this course, you should expect to present one or two research papers during the course. The instructor will ask students to sign up papers to present by the end of the first week. The instructor will present the rest of the unselected papers during the course.

Each paper presentation should be no more than 30 minutes, so we can have enough time for discussion. The presentation should focus on elaborating the motivation, related work, tool/study design, research questions, findings, limitations, and future work of the assigned paper. To make your presentation more insightful, try to center your presentation based on the literature and tell the audience why this work is proposed in the first place, how it advances people's understanding about a topic, and how it is different from other related work in the past. You are also encouraged to connect the assigned paper to your own research. You should prepare for a set of questions (either came up by yourself or based on questions other students post on Piazza) and co-lead an in-class discussion with the instructor based on these questions after the presentation.

The in-class discussion will follow the think-pair-share format.

- 1) Think. The presentor or the instructor will provoke students' thinking with a question. The students should take one or two minutes just to THINK about the question.
- 2) Pair. Using designated partners (such as with Clock Buddies), nearby neighbors, or a deskmate, students PAIR up to talk about their answers with each other. They compare their mental or written notes and identify the answers they think are best, most convincing, or most unique.
- 3) Share. After students talk in pairs for a few minutes, the presenter or instructor will call for pairs to SHARE their thinking with the rest of the class.

Course Project Instructions

You are expected to work on a course project either alone or in groups. You can pick any topics related to human-AI interaction. In Week 2, I will release a few sample project ideas to guide you with the process of choosing a project topic. Between Week 2 and Week 5, please stop by during office hours to discuss your project ideas with the instructor to get early feedback on the novelty, feasibility, and significance of your ideas.

A short project proposal is due on Week 5. This proposal should describe the project idea, the motivation of this idea, and (optional) a usage scenario if you propose to build a new tool. The proposal could be any length but no longer than 4 pages. It will be evaluated based on the quality of the idea and writing, not the length of your writing.

A mid-point project summary is due on Week 9. This summary should describe the envisioned approach/methodology/design as well as which parts have been done so far. The summary could be any length but no longer than 4 pages.

In Week 15, each team will deliver a presentation of their project. The presentation will be about 20 minutes. You will get another 5 minutes for Q&A.

A final project report is due on the final exam week (max 10 pages plus references). Your final project report should be built upon your proposal and project summary. Feel free to reuse sections from those two reports in your final report. You may include an appendix beyond 10 pages, but your paper must be understandable without it. Submissions should be in the ACM format.

Your final report should be structured like a conference paper. It should contain:

- Abstract
- A well-motivated introduction
- Related work with proper citations
- Description of your methodology
- Evaluation results
- Discussion of your approach, threats to validity, and additional experiments
- Conclusions and future work

If you are doing a project that involves implementation, please include a link to your Github repository in your final report. Please also add a README file in your repository to describe how to run and test your code.

Quizzes

We will have five quizzes during this course. Each quiz will assess your understanding about the research topics we have covered in the previous two or three weeks. The quizzes will include multiple-choice questions and open-ended questions. To prepare for the quizzes, make sure (1) you have read the papers, (2) review the slides from the instructor or the paper presenter, (3) understand the methodologies, findings, and contributions of each paper. The instructor cannot accommodate quizzes on a different date. However, we will count the best 4 quizzes out of 5 quizzes. Each quiz should take about 20 minutes.

Course Policies and Expectations

Attendance

Please come to the class continuously, read the assigned papers, and participate into discussions. While we will not check attendance in each class, we will use other ways including quizzes, we will have five quizzes. Your final grade will also depend on your participation in the class. So please come to the class continuously and participate in all required activities.

Late submissions

Late submissions of assigned work will be accepted with 7.5% decaying credit per day.

Time commitment

Students are expected to spend no more than 12 hours per week in class or on coursework per week on average. I suggest, early in the semester, setting aside 3 hours to read and complete any assigned work related to the assigned papers before each class, leaving approximately 3 hours per week for group formation, early scouting of research project topics, need-finding research and brainstorming. As we get deeper into the semester, I suggest spending 2 hours per class reading assigned papers and completing related assigned work, leaving approximately 5 hours for building prototypes, conducting user studies, and preparing presentations.

Feedback to the instructor

During this course, I will be asking you to give me feedback on your learning in both informal and formal ways. Occasionally, at the end of a lecture, I will hand out index cards to collect anonymous comments and questions about this class and your learning experience. In the middle of the semester, I will send out an anonymous midpoint survey about how my teaching strategies are helping or hindering your learning. It is very important for me to know your reaction to what we are doing in the class, so I encourage you to respond to these surveys, ensuring that we can create an environment effective for teaching and learning.

Grading

Reading assignments [20%]

Paper presentation [20%]
Final project [40%]
Quizzes [10%]
Class participation and discussion [10%]

Academic Integrity

Discussion and the exchange of ideas are essential to academic work. For assignments in this course, you are encouraged to consult with your classmates on the choice of paper topics and to share sources. You may find it useful to discuss your chosen topic with your peers or course instructional staff, particularly if you are working on the same topic as a classmate. However, you should ensure that any written work you submit for evaluation is the result of your own research and writing and that it reflects your own approach to the topic. You must also adhere to standard citation practices in this discipline and properly cite any books, articles, websites, lectures, etc. that have helped you with your work.