MUTUAL THEORY OF MIND FOR HUMAN-AI COMMUNICATION IN AI-MEDIATED SOCIAL INTERACTION

A Dissertation
Presented to
The Academic Faculty

By

Qiaosi Wang

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy in the
School of Interactive Computing
Human-Centered Computing

Georgia Institute of Technology

December 2024

MUTUAL THEORY OF MIND FOR HUMAN-AI COMMUNICATION IN AI-MEDIATED SOCIAL INTERACTION

Thesis committee:

Dr. Ashok K. Goel (Advisor)
School of Interactive Computing
Georgia Institute of Technology

Dr. Lauren G. Wilcox
Responsible AI
eBay & Georgia Institute of Technology

Dr. Elizabeth N. Disalvo
School of Interactive Computing
Georgia Institute of Technology

Dr. Q. Vera Liao
FATE Group
Microsoft Research, Montreal

Dr. Munmun De Choudhury
School of Interactive Computing
Georgia Institute of Technology

Date approved: September 12, 2024

To my parents and Jerry.

ACKNOWLEDGMENTS

Before I started my PhD program, I thought of it as an opportunity to learn everything about research. But it turned out to be so much more of a journey to learn about who I am and who I want to become as a researcher. Along the way, I've been incredibly fortunate to meet and work with some of the most supportive, kind, hard-working, and brilliant individuals I've ever known. They have not only shaped my approach to research but also played a significant role in shaping the person I am today.

I want to first thank my advisor Ashok Goel for being the best advisor I could've ever asked for. When I chose to work with Ashok, I never anticipated that Ashok would become the embodiment of a perfect PhD advisor— someone that I aspire to be for my own students if I ended up pursuing a career in academia. Throughout my PhD training, Ashok has always treated me as his equal, given me endless freedom, support, patience, trust, and encouragement I needed to confidently pursue my ideas, no matter how unpolished those ideas initially were. Ashok was instrumental in guiding me through the development of Mutual Theory of Mind, providing invaluable expertise and insights throughout the process of writing this thesis——something that would not have been possible without his support.

I am also grateful for the guidance and feedback from my committee members. Lauren Wilcox has been my strongest advocate since day one of my PhD program. Lauren has supported me and offered me valuable opportunities at critical turning points, and I have always found comfort knowing she is rooting for me. Betsy Disalvo was among the first to recognize my research potential and instill confidence when I needed it the most. I will miss the times when I would "casually" run into Betsy just outside of her office during periods of research anxiety and she would always greet me with her calming presence. Vera Liao's work has been a huge inspiration to me throughout my PhD training. I first reached out to Vera at CSCW 2020 as a junior PhD student and quickly made her my role model after our brief 30-min chat— I hope some day I can match Vera's kindness, enthusiasm, sharpness,

and curiosity in research. Munmun De Choudhury's work, perspectives, and methodology have been deeply influential in shaping my PhD research. Her feedback was invaluable in refining my thesis, and I will always hold the utmost respect for her.

It truly takes a village to raise a PhD student. I was really lucky to have worked with some of the best minds in HCI during my internships and various collaborations. Many of them became my mentors that have guided me through some of the most stressful times. Michael Terry, Michael Madaio, and Michael Muller provided much guidance and support while I was navigating the job market and my future research career in early 2024. I am grateful to still be able to count on their wisdom and advice from time to time. I met Justin Weisz through our shared interest in Mutual Theory of Mind and I have had the honor to work on several MToM efforts with him since then. I have learned so much from Justin's passion, dedication, and rigorous attitude and I hope to continue our collaboration for many more years to come. David Joyner was one of my first collaborators and my academic brother. David has provided research and emotional support from my first first-author paper draft to my thesis defense, and I have benefited much from his wisdom, humor, and kindness. I also want to thank Jeff Kephart, Mei Si, Thomas Ploetz, and Shaun Kane who have provided valuable feedback during our collaborations and have made me a better researcher in various ways. I wouldn't have pursued a Ph.D. in HCI without the guidance and inspiration of my mentors during my undergraduate studies at the University of Washington. Christina Chung, Sean Munson, and Julie Kientz opened up the door of HCI research for me and were kind and patient enough to mentor an undergraduate student to do HCI research out of their insanely busy schedules. Laura Little taught me the fundamentals of statistics and experimental design, while also showed me the remarkable extent to which a professor can support their students.

One of the most valuable advice I have given to junior students is that doing a PhD is lonely, but it doesn't have to be. I want to thank my amazing friends for helping me realize that. Vedant Das Swain has been my go-to sounding board for random research ideas,

a judgment-free space for my basic quantitative method questions, and one of my closest friends who supported me through the toughest time of my PhD journey. Jiawei Zhou has tolerated my annoyance and tantrums, and showed me unwavering support, kindness and friendship. Dong Whi Yoo has been a major source of comfort, wisdom, and encouragement ever since my initial days in the PhD program. I also want to thank my friends Sungeun An, Koustuv Saha, Yu Fu, Sarah Walsh, India Irish, Karthik Bhat, Rui Zhou, and Karan Taneja, who have supported me throughout various periods of my PhD and we had so much fun together during spontaneous Starbucks runs, coffee chats, and various celebrations throughout the years. Diana Qu, Fanlu Gui, Christine Hao, and Vikram Mohanty have witnessed my PhD journey from the side yet still provided endless support, empathy, and encouragement that got me through my PhD program. I also want to thank the many students and research scientists I collaborated with at DILab, who have inspired me to be a better mentor and collaborator: Shan Jing, Ida Camacho, Eric Gregori, Sandeep Kakar, Chidimma Anyi, Jingying Zeng, Benjamin Faught, Chris Leung, Rhea Basappa, and Lingqing Wang.

This year marks the 11th year since I left home to pursue further education in the United States. My parents have made many sacrifices and broken many societal stereotypes to send their little girl to study overseas to get her bachelor degrees, and now a Ph.D. degree. I hope this PhD thesis makes them proud and provides some solace for the time I wasn't able to spend by their side. Throughout the six years of my PhD training, my husband Jerry Li has tolerated my frequent yet random burst of anxiety, taken care of chores during the weeks leading up to the CHI deadlines without any complaints, consoled me during my low points day and night, and moved across the country to accompany me. Hopefully this thesis provided greater meaning to your sacrifices in the past six years and I can't wait to spend more quality time with you and our cat Gouda from now on.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ACRONYMS

AI Artificial Intelligence

CA Conversational Agent

CSCW Computer-Supported Cooperative Work

HAI Human-AI Interaction

HCI Human-Computer Interaction

JW Jill Watson

MToM Mutual Theory of Mind

OMSCS Online Master of Science in Computer Science

QA Question-Answering

RTA Reflexive Thematic Analysis

SAMI Social Agent Mediated Interaction

ToM Theory of Mind

SUMMARY

AI systems are being quipped with Theory of Mind (ToM)-like capability to advance their social adeptness while they assume diverse social roles in our society. ToM is a basic social and cognitive human capability of attributing mental states such as beliefs, emotions, knowledge, and plans to oneself and others based on behavioral or verbal cues. AI systems with ToM-like capability can infer humans' mental states and customize their responses to cater to our knowledge, needs, goals, preferences, etc.. As these AI systems exhibit human-like "mind-reading" capability to enhance human-AI communications and collaborations, people are increasingly uncertain about how they should perceive such AI systems' social roles and capabilities. Mismatch between AI systems' actual social capabilities and people's expectations of the AI can not only lead to user frustrations and abandonment of the AI systems, but also lead to harms such as overreliance on AI and extreme anthropomorphism of the AI. Managing and accounting for human perceptions of AI systems performing at varying social capacities becomes crucial in improving user experience and mitigating harms in human-AI communications.

Inspired by the Mutual Theory of Mind in human-human social communication, this thesis posits the MToM framework for human-AI communication, in which both the humans and the AI systems can leverage their ToM-like capability to continuously construct, recognize, and respond to others' perceptions of them. The MToM framework aims to guide the research and design of human-AI communication by breaking down the iterative human-AI communication into three analyzable stages. Each MToM stage represents a ToM process of one party's communication feedback shaping the other's interpretation of how they are perceived by others. As a first step towards realizing the vision of MToM in human-AI communication, this thesis followed the MToM framework and conducted a series of empirical studies to provide design implications for building AI systems' ToM-like capability to account for human perceptions of AI during communications. Each study

examines human perceptions of AI at a MToM stage: ToM construction (AI's construction of human's interpretation), ToM recognition (human's recognition of AI's interpretation), and ToM revision (AI's revision of its interpretation). These studies were conducted in the context of AI-mediated social interaction in large-scale learning environments, where AI systems are already leveraging their ToM-like capability to provide personalized social recommendations to adult learners based on information inferred from their digital footprints.

To understand the social role and desired capabilities of AI systems in AI-mediated social interaction in large-scale learning, this thesis first empirically explored the design requirements of such AI systems from online learners' perspectives. Through two studies with online learners, I pointed out the design opportunities for AI systems to perform the role of social facilitators while enhancing the social translucence and bridging the social-technical gap in remote social interaction. I further outlined implications for designing human-like AI agents that can alleviate challenges in remote social connections, cater to students' evolving social needs, and mitigate potential privacy and social harms in AI-mediated social interaction. These findings established the need for AI agents to account for students' perceptions of the AI during AI-mediated social interaction in online learning.

The rest of the thesis empirically explored students' perceptions of AI in AI-mediated social interaction at each stage of the MToM framework in large-scale learning. At the ToM construction stage, I conducted a longitudinal survey study that highlighted students' evolving perceptions of the AI agent over time, and established the feasibility for AI agents to automatically construct students' perceptions of AI through analyzing linguistic characteristics of students' utterances to the AI. At the ToM recognition stage, I conducted a mixed-methods study to understand students' reactions and perceptions of AI after recognizing AI's (mis)interpretations of their personal characteristics in AI-mediated social interaction. I identified three types of rationales that students adopted to make sense of the AI's misinterpretations of their personal characteristics. I found that these rationales are informed by students' evolving AI knowledge through AI output, and can lead to harmful perceptions

and reactions to AI misinterpretations. At the ToM revision stage, I conducted a mixed-factorial vignette experiment to examine the impact of AI's revision of its misinterpretations on students' perceptions of the AI. I found that AI systems can mimic human's ToM revision reasoning and communication process to mitigate students' negative perceptions of AI after encountering AI misinterpretations. However, balancing the informational and social aspect of AI's revision communication is the key to mitigate students' perceptions of the AI.

Overall, this dissertation makes theoretical, design, and empirical contributions to the fields of human-AI interaction, computer-supported cooperative work, and responsible AI. This work provides theoretical guidance and actionable design implications to build the next generation of AI systems that can continuously construct, recognize, and respond to human perceptions of AI in human-AI communication.

CHAPTER 1
INTRODUCTION

1.1 Motivation

New developments in Artificial Intelligence (AI) have enabled AI systems to assume diverse social roles as our assistants and partners in human society. To facilitate their social functions, AI systems have been designed with human-like physical appearances, anthropomorphic expressions and behaviors, and natural language communication abilities. More recently, AI systems are being equipped with Theory-of-Mind-like capabilities to further advance their social adeptness. Theory of Mind (ToM) is a basic social and cognitive human capability of attributing mental states such as beliefs, emotions, knowledge, and plans to oneself and others based on explicit or latent behavioral and verbal cues [1, 2, 3, 4, 5, 6]. AI systems with ToM-like capability [1] can make inferences about human's mental states and customize their responses to cater to our knowledge, needs, preferences, etc. [12, 13, 14, 15, 16, 17]. Such AI systems present great promises in enhancing the efficiency, naturalness, and human experiences across varying human-AI interactions. For instance, in human-AI task collaborations, AI systems with ToM-like capabilities can detect human collaborators' knowledge state and provide additional information accordingly to complete the tasks [18], infer human intentions and plans to account for irrational human behaviors during collaborations [19], and generate legible robot motions based on predicted human interpretations [20]. In the near future, AI with ToM-like capabilities can be applied in more aspects of our lives: imagine an AI agent that can facilitate and promote social con-

[1] Whether AI systems can or will possess ToM is a highly debatable topic in current academic discourse [7, 8, 9, 10, 11]. This thesis is based on my current belief that AI presently does not have human-level ToM capability and can only possess ToM-like capabilities to generate human-like behaviors. Therefore, throughout this thesis, I use "ToM-like capabilities" to refer to AI's capability of inferring and attributing mental states to people with the goal of predicting human behaviors.

nections among fully remote workers or online students, an AI teaching assistant that can detect students' confusion about class materials and offer personalized guidance, or an AI assistant that can infer and protect our focus work time against outside distractions [21].

As such AI systems continue to exhibit seemingly human-level ToM capability, people are becoming more uncertain than ever about how they should perceive AI systems' roles and capabilities— people are expecting these human-like machines to perform social functions at the human level. While failure in matching those expectations can lead to user frustrations or even abandonment of the AI [22, 23], performing beyond those social expectations can lead to greater harms such as overreliance on AI [24, 25, 26, 27], self-disclosure of sensitive information to AI [24, 28, 27, 29], and extreme anthropomorphism such as viewing AI as romantic partners or mental health therapists [24, 28, 27]. **Managing and accounting for people's perceptions and expectations of AI systems performing at various social capacities becomes a critical problem for improving user experience and mitigating potential harms in human-AI interaction.**

In human-human communication, we are able to leverage our ToM capability to actively engage in the ToM process of constantly inferring about others' perceptions of us through social cues embedded in their behavioral and verbal feedback. This is the *ToM process* of constructing one's theory of the other's mind. Goffman(1978)'s seminal work on impression management further suggests that people can not only infer others' perceptions of them, but also leverage various social techniques to intentionally shape others' perceptions of them through behavioral and verbal feedback. In social communications, inferring one specific dimension of others' minds— their perceptions of us— can help us behave accordingly to match or shape others' perceptions of us [30]. When both parties in the communication leverage their ToM capability in this way, which I call "Mutual Theory of Mind (MToM)", can enable them to continuously engage in the iterative ToM process of constructing, revising, and responding to others' perceptions of them to maintain proper social expectations of each other, which ensures smooth and continuous communication

and avoids social awkwardness [30]. However, current human-AI communication has not achieved such MToM. Without AI's capability of recognizing human's perceptions of AI and provide cues to shape people's perceptions, human's perceptions of AI often remain uninformed. This has left the burden of communication to the humans to figure out how they should accurately perceive the AI's social roles and capabilities through endless trial-and-error [22].

Inspired by the MToM in human-human communication, this dissertation envisions human-AI communication through the lens of MToM, where both the human and the AI system can use their ToM-like capability to construct, recognize, and respond to how they are interpreted by the other party, therefore resulting in accurate interpretations of each other's roles and capabilities. **To achieve this vision of MToM, this dissertation posits the MToM framework with the goal of guiding the research and design of MToM in human-AI communication.** The MToM framework breaks down the iterative human-AI communication process into three analyzable stages: ToM construction, ToM recognition, and ToM revision. Each stage demonstrates the ToM process of one party's feedback shaping the other party's interpretation of how they are perceived. **Guided by this framework, this thesis then empirically examines human perceptions of AI systems that assumed the social role of match-makers, aiming to offer design implications on building AI systems' ToM-like capability that can account for human perceptions of AI at every MToM stage in human-AI communication.** Specifically, I took a mixed-methods approach to study the feasibility of AI's automatic *construction* of human perceptions of AI, human's perceptions and reactions to the AI after *recognizing* AI's (mis)interpretations of humans' personal characteristics, and the design of AI's *revision* of its misinterpretations that can shape human perceptions of AI. **This thesis contextualizes MToM for human-AI communication in the context of AI-mediated social interaction in large-scale learning environments in higher education.** These AI systems are increasingly common in large-scale learning environments to enhance students' learning experiences. AI systems

performing AI-mediated social interaction are often equipped with ToM-like capability to make personalized social recommendations to students by inferring students' mental states (e.g., emotions [15], preferences [31]), and characteristics (e.g., personalities [32, 33]), with the goal of improving students' social connectedness and learning outcomes [34, 35, 36]. This makes AI-mediated social interaction in large-scale learning environment a great context to study the social communications between humans and AIs equipped with ToM-like capability.

In the rest of this chapter, I first describe in detail about the MToM framework for human-AI communication by defining ToM in the MToM framework, outlining the three stages of the communication process, and describing the *ToM process* happening at each stage by highlighting the three critical elements involved. I then apply the MToM framework to the context of AI-mediated social interaction in large-scale learning environment to identify research questions in this space to enhance human-AI communication through the lens of MToM. Next, I provide an overview of the thesis context of human-AI communication in large-scale learning environments. In Thesis Overview, I motivate my research questions in this thesis, summarize my thesis studies, introduce my thesis statement, and describe my thesis contribution. I conclude this chapter by presenting an organization of this dissertation.

1.2 Thesis Framework

1.2.1 The Mutual Theory of Mind Framework for Human-AI Communication

Inspired by the MToM in human-human social communications, the MToM framework (as shown in Figure 1.1) for human-AI social communication breaks down the iterative communication process into *three analyzable stages: ToM construction, ToM recognition, and ToM revision*. Given that communication is a two-way street where both parties' interpretations of each other are constantly shaped by the other's communication feedback, the MToM framework captures this iterative process by assigning specific roles to each

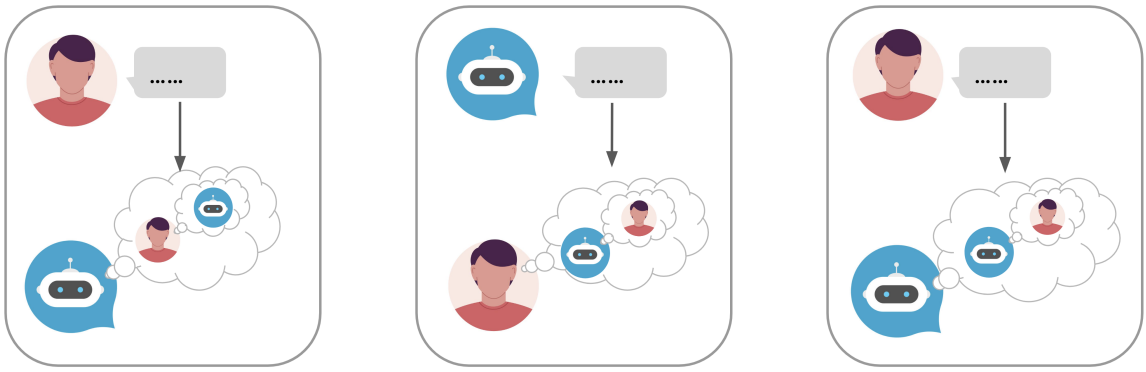


Figure 1.1: The Mutual Theory of Mind framework breaks down the human-AI social communication process into three stages: ToM construction, ToM recognition, and ToM revision. Each stage illustrates the Theory of Mind process of how one's **feedback** (represented as rectangular bubble) can **shape** (represented as arrow with directionality) the other's **interpretation** (represented as cloud bubble) of how they are perceived. This figure outlined the MToM framework with a basic role assignment and communication order modeled after a typical human-initiated, regular turn-taking human-AI communication process.

party at every MToM stage: ToM construction refers to the process of AI's construction of human's interpretation of the AI based on human feedback; ToM recognition refers to the process of human's recognition of AI's interpretation of the human based on AI feedback; and ToM revision refers to the process of the AI's revision of its prior interpretation of the human based on human feedback. Each stage describes the ToM process by highlighting the interaction between the *three elements: interpretation, feedback, and shaping*.

Defining "Theory of Mind" in the MToM Framework. In the MToM framework, ToM refers to the *process* of one's *interpretation* of each other being iteratively *shaped* by communication *feedback*. This is in contrast with much existing literature that have studied ToM as a capability of making inferences about others' mental states [9], or as a representation of others' mental states [37]. Inspired by impression management theory [30], such interpretations during social communications are often nested, i.e., *my interpretation of your interpretation of me*, and can be actively shaped by communication feedback from the other

party. This type of nested interpretation, similar to inferences about self or the other party's emotions, beliefs, or knowledge, is one specific dimension of one's mental state that can be shaped and inferred through communication feedback, and forming a MToM in human-AI communication. Given that mental states can be attributed to self as well as others in the ToM process, such nested interpretations can present as "my interpretation of your interpretations of me" or "my interpretation of my interpretation of you." As illustrated by the MToM framework, the subjects and objects of such interpretations are interchangeable at different stages, e.g., AI's construction of human's interpretation vs. human's recognition of AI's interpretation, highlighting the iterative nature and mutual shaping process of human-AI communication.

Three Elements of the ToM Process in Human-AI Social Communication. In human-human social communication, one can use their ToM capability to form an interpretation of how they are perceived by the other party based on others' communication feedback. This process therefore consists of three elements: interpretation, feedback, and shaping. Throughout the human-AI social communication process, humans and AIs can each construct, derive, and revise their *interpretation* such as how they interpret the other party's interpretation of them (e.g., ToM construction stage), or how they interpret their interpretation of the other party (e.g., ToM revision stage) based on the other's communication feedback. As mentioned in the previous section, such interpretation in the MToM framework are often nested to maintain proper social expectations and impressions. *Feedback*, often in the form of verbal (e.g., text messages) or behavioral feedback (e.g., gestures), can be generated with different complexities based on one's interpretation to either match or shape the other party's impression of them. When both the human and the AI are equipped with ToM-like capability to facilitate social communication, they can employ such capability to generate feedback that can actively *shape* the other party's interpretation.

Three Stages of the MToM Framework. Through the lens of MToM, the continuous and iterative human-AI social communication process can be broken down into three stages: ToM construction, ToM recognition, and ToM revision. Each stage describes a ToM process of feedback shaping interpretation. Modeled after a typical human-initiated human-AI communication process, the first stage, *ToM construction*, describes the process of the human providing communication feedback to the AI, which was analyzed by the AI system to *construct* human's perceptions of the AI. In the second stage, the AI system provides communication feedback to the human that conveys its interpretation of the human, which could be *recognized* by the human. For example, humans can typically scroll through the list of products recommended to them by the AI system and recognize the AI's interpretations of their preferences. In the third stage, humans can provide feedback to the AI system to improve the accuracy of the AI's feedback. The AI can then *revise* its interpretation of the human by incorporating the human feedback and introspecting on what led to the inaccurate AI feedback in the first place. After revising its interpretation, the AI system can communicate its revision to the human to update and shape the human's interpretation of the AI.

It should be noted that while the current MToM framework is modeled after a typical human-initiated human-AI communication process with regular turn-taking, *the MToM framework acknowledges that human-AI communication can take different forms and processes and embrace variations of the framework.* Depending on the specific communication context, researchers have the flexibility of assigning different roles to each communication party as well as varying orders of the three communication stage to study different human-AI communication. For example, ToM construction stage can also describe the human's construction of AI's interpretation of the human; the order of the human-AI communication process can also start with ToM recognition instead of ToM construction. In this thesis, I use the MToM framework with the basic role assignment and communication order outlined in Figure 1.1.

The MToM framework provides guidance on how to design towards Mutual Theory of Mind in human-AI social communication, where both the humans and the AIs can constantly infer and respond to others' impressions of them. The MToM framework can be applied to a variety of context where AI systems assume diverse social roles to communicate with humans. This thesis specifically applies the MToM framework to the context of AI-mediated social interaction in large-scale learning environment, where AI systems are playing the role of social match-makers to facilitate students' social connection process.

1.2.2 Applying the MToM Framework to AI-Mediated Social Interaction

AI systems that can mediate students' social interaction process are often equipped with the capability of inferring about students' characteristics to provide personalized social recommendations. In this process, the AI system takes up the role of a social match-maker to connect students with shared hobbies, cities, interests, classes, or career goals together to provide social support throughout the education program. In this thesis, I apply the MToM framework to AI-mediated social interaction to empirically examine the design requirements of such AI system towards MToM in human-AI communication. I began by envisioning MToM in such human-AI social communication through an example dialogue between a student and an AI agent during AI-mediated social interaction, as shown in Figure 1.2.

In this dialogue, the student initiates the conversation with the AI agent by asking about the agent's capability, while conveying uncertainty about the AI agent's role in making the student feel socially connected. Based on the student's message, the AI agent can *construct* the student's perception of the AI then understand that the student might have formed unrealistic expectations and perceptions of the agent (e.g., treating the agent as their friend). Based on this interpretation, the AI agent should be able to respond accordingly to correct student's perception of the AI and provide social recommendations to the student based on

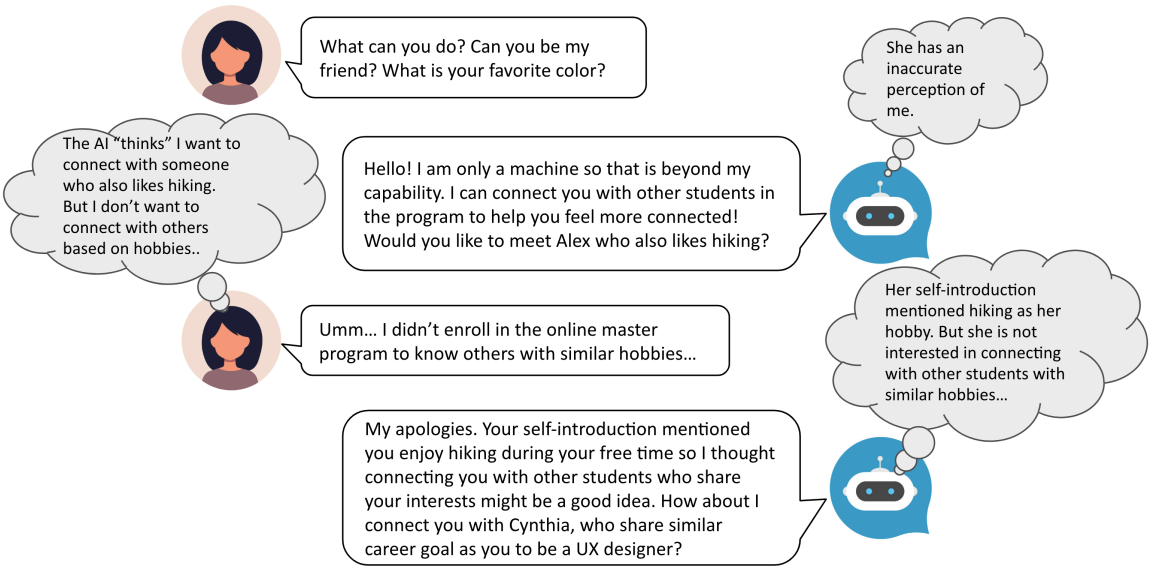


Figure 1.2: An example AI-mediated social interaction dialogue in an online learning environment through the lens of MToM. This dialogue illustrates how the student and the AI agent's feedback can shape each other's interpretations throughout the three stages: ToM construction, ToM recognition, and ToM revision.

the student's data that the agent has access to. From the AI's social recommendation, the student can *recognize* the AI's interpretation of their social needs and preferences, which might be inaccurate. In order for the AI agent to provide more accurate social recommendation, the student can respond with information to enhance the AI's interpretation of their social preferences. The AI agent can take in the student's feedback and *revise* its previously inaccurate interpretation about the students' social preferences. And finally, respond with its revision of the inaccurate interpretation to the student.

As illustrated, by applying the MToM framework to the context of AI-mediated social interaction in large-scale learning environment, I was able to envision a human-AI social communication process where the human and the AI agent can constantly and iteratively construct, recognize, and respond to their interpretations of each other for natural, continuous, and effective human-AI communication. While humans already possess such ToM capability to make conjectures about the AI's interpretation of the human and respond with feedback to correct the AI's interpretation, AI systems still lack such ToM-like capability

to construct human's perceptions of the AI and respond accordingly to facilitate the interaction as illustrated in Figure 1.2. To realize this vision, this thesis offers implications on designing such AI system from a human-centered perspective by examining students' perceptions of AI at each stage of this communication process in AI-mediated social interaction in the context of large-scale learning environment.

1.3 Thesis Context

Large-scale learning environments have been widely adopted by higher education in the form of large in-person classrooms or online for-degree programs that can fulfill adult learners' increasing needs for lifelong learning and workforce development [38]. However, large-scale learning environments often sacrifice individual needs for scale— in in-person or online classrooms with hundreds or thousands of students, it is difficult for instructors to identify individual students' learning needs and goals; it is equally difficult for students to collaborate, discuss, and connect with their fellow classmates [39]. To address these issues, higher education has adopted AI systems, especially anthropomorphized AI agents, to provide students with personalized support. These AI systems are assuming diverse social roles in the classrooms, such as teaching assistants [40], social facilitators [36], and writing partners [41] to provide personalized support to individual students. To better fulfill students' individual needs, these AI systems are being increasingly equipped with varying levels of ToM-like capability to infer about students' mental states such as emotions, knowledge, needs, and goals from their digital footprints in large-scale learning environments (e.g., class discussion forum posts, assignments submitted). However, students' perspectives and interpretations of these AI systems equipped with ToM-like capabilities remain under explored. Examining students' perceptions and interactions with these AI agents can provide critical implications that can improve the usability and effectiveness of such AI agents to properly respond and fulfill students' individual educational needs and goals in the program. These implications could also provide insights into the transferability of the MToM

framework when applied to other human-AI communication contexts, given adult learners' diverse AI literacy level and cultural and demographic backgrounds.

With these goals in mind, this thesis mainly examines human-AI communications between adult learners and AI agents that act as social facilitators to perform AI-mediated social interaction in large-scale learning contexts. These AI agents leverage NLP techniques to extract information from adult learners' digital footprint, such as their demographic information and career goals from students' self-introductions paragraphs, typically posted on the class discussion forum [42]. Based on information extracted, these AI agents provide personalized social recommendations to facilitate social interactions among adult learners, including team-matching. The ultimate vision for such AI social facilitators is that they can make inferences about students' career goals, academic interests, and other implicit educational needs and goals to connect like-minded students, and therefore enhance social presence in large-scale learning contexts [39, 43]. Much of this thesis (chapter 4, chapter 6, chapter 7) examines the design of a specific AI agent named SAMI (stands for "Social Agent Mediated Interaction") [36] that can perform social match-making among students based on inferences from students' self-introductions (details about SAMI can be found in chapter 3).

This thesis begins by taking the Online Master of Science in Computer Science (OMSCS) program at Georgia Tech as an exemplar to examine online adult learners' perspectives and interactions with AI agents taking social roles in the classroom in chapter 4 and chapter 5. To better fulfill students' individual learning needs and goals, the OMSCS program is dedicated to incorporate AI agents to support students. The program has been a testbed for AI agents that are playing social roles like teaching assistants and social facilitators for quite some time [40, 44, 36]. The OMSCS program was established in 2014 and have graduated more than 10,000 adult learners in its first 10 years. The OMSCS program consists of students from all over the world that are eager to learn and earn a computer science master degree. The program currently offers more than 50 computer science classes in

an asynchronous format, with 200-1000 students enrolled in each class. Students often take these classes by using a combination of communication technology such as Canvas, online class discussion forums, and Slack channels, all of which present opportunities to incorporate AI agents to enhance students' learning experiences. Many students are attending the program part-time while working full-time jobs ($\tilde{8}7\%$), with an average program starting age of 30 years old. More details about the program and student demographic can be found in chapter 3.

Based on the findings of online students' perspectives and communications with AI agents in the OMSCS program, the other studies presented in chapter 6 and chapter 7 further examined the perspectives from students in large-scale learning contexts including large in-person classrooms as well as students from more diverse study majors, i.e., non-STEM majors. These students were recruited either from Georgia Tech or the Prolific crowd-sourcing platforms in the U.S.. In general, most of these students were at the undergraduate level, with about half of the students studying in non-STEM major such as business, nursing, psychology, etc. These students have an average age of 30 ± 10 and ranges between 18 to 73 years old. The details about the participant demographic and background can be found in each study chapter, chapter 6 and chapter 7.

1.4 Thesis Overview

The goal of this thesis is to *offer guidance and implications on how to design AI systems assuming social roles that can account for human perceptions of AI to realize the vision of Mutual Theory of Mind in human-AI communication.* To achieve this goal, this thesis proposed and applied the MToM framework to examine students' perceptions of AI agents assuming social roles at each stage of the MToM process in large-scale learning environment. I began this exploration by understanding students' perspectives on designing AI systems with ToM-like capability to perform AI-mediated social interaction by answering this question, with the OMSCS program as an exemplar of large-scale learning context:

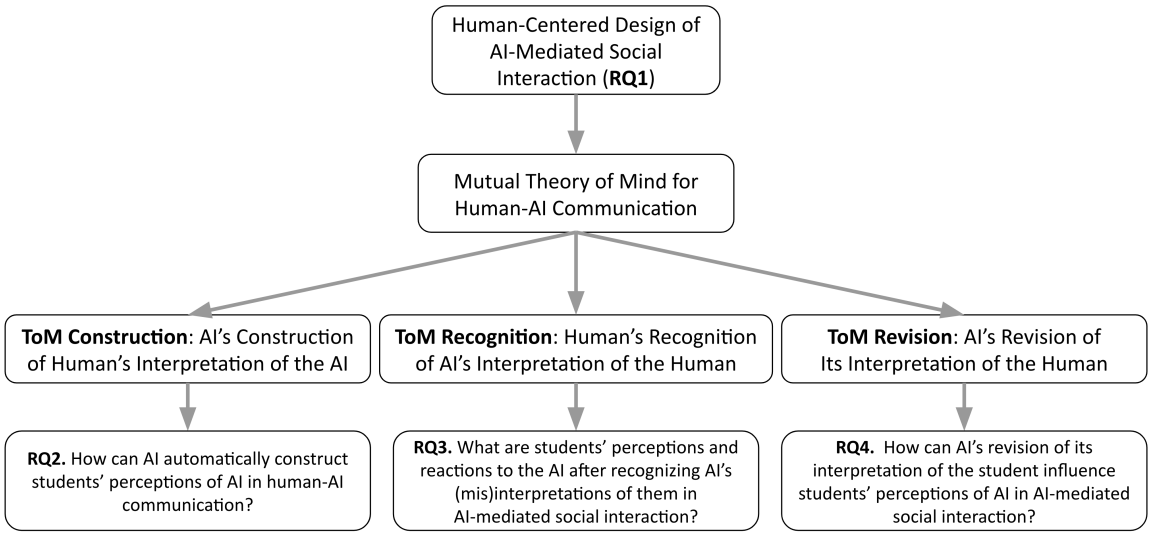


Figure 1.3: Thesis overview.

RQ1. What are the design requirements of AI-mediated social interaction from online learners' perspectives?

I explored this question through two studies. First, I conducted semi-structured interviews with online learners to understand their existing challenges, needs, and practices in building social connections remotely. Building upon the findings from this study, I conducted a series of co-design workshops with online learners and distilled a set of design guidelines that detail the desired functionalities, social characteristics, and ethical concerns of AI agents that can perform AI-mediated social interaction in online learning. These two studies suggested online learners' preferences for human-like AI agents to mediate online learners' social interaction process through continuous communications to cater to students' evolving social needs and goals, as well as to alleviate the challenges in remote social connections. However, I also uncovered potential privacy and social risks due to students' harmful perceptions of the AI. These two studies established the need to design AI agents with ToM-like capability that can account for students' perceptions of the AI during AI-mediated social interaction.

Inspired by the MToM in human-human communications where both parties continu-

ously construct, recognize, and revise their interpretations of each other, I proposed the MToM framework to envision MToM in human-AI communication and outline the critical processes and elements to achieve the vision of MToM. The rest of this thesis applies the MToM framework in AI-mediated social interaction to gain design implications of how to design AI system's ToM-like capability to account for human perceptions of AI throughout each MToM stage. The first stage of MToM in human-AI communication is for the AI system to *construct* students' perceptions of the AI based on students' communication feedback. This will equip the AI system with the basic ToM-like capability to monitor student's changing perceptions and provide communication feedback that can help the students build a better mental model of the AI. Specifically, I examined this question:

RQ2. How can AI automatically construct students' perceptions of AI in human-AI communication?

Through longitudinal surveys and linguistic analysis, I established the feasibility for AI systems to automatically construct students' perception of the AI by analyzing the linguistic characteristics of students' utterances to the AI. This study also pointed out that students' perceptions of the AI agent can fluctuate over time even when the AI agent does not have learning capability to improve its performance over time, highlighting the importance of continuous monitoring of student's perceptions for the AI agent to respond accordingly.

Following the MToM framework, I then explored students' perceptions of AI at the ToM recognition stage. Studies from the earlier chapters identified students' preferences for AI agents' social functionalities and characteristics, and surfaced some potential ethical issues stemmed from students' various perceptions of AI agents, e.g., overtrusting AI agents serving social purposes (chapter 4). To better understand students' perceptions and reactions to AI's social behaviors, in this chapter, I examined students' perceptions of AI after recognizing AI's misinterpretations of students' personal characteristics in AI-mediated social interaction. AI systems that can profile people's personal characteristics such as personalities [33, 45] can give people the illusion that "machines can read our minds." [46]

This illusion has led to various rather concerning reactions and perceptions of AI such as attributing AI systems with beyond-human expertise at reading people's emotions and personalities [45, 47]. However, people's perceptions and reactions of such AI systems when this illusion is broken in the face of AI misinterpretation have not yet been explored. Understanding people's reactions and perceptions of the AI after encountering AI misinterpretations can provide critical implications for the future design and development of responsible interventions, mitigation, and repair strategies to retain user trust, minimize harms, and prevent overreliance when such AI systems inevitably err. I focus my exploration of the ToM recognition stage by answering this question:

RQ3. What are students' perceptions and reactions to the AI after recognizing AI's (mis)interpretations of them in AI-mediated social interaction?

To answer this question, I took a Wizard-of-Oz approach to fabricate intentionally inaccurate/accurate AI interpretation of students' personalities to elicit their perceptions and reactions to the AI. Through semi-structured interviews and a large online survey experiment with students, I found that students' existing and newly acquired AI knowledge plays a critical role in shaping their perceptions and reactions after encountering AI (mis)interpretations of their characteristics. Specifically, I pinpointed three rationales that students adopted through knowledge acquired from AI (mis)interpretations: AI works like a machine, human, and/or magic. These rationales can determine their perceptions and reactions to the AI after recognizing AI's interpretations of them. Certain rationales can lead to dangerous perceptions and reactions to the AI, such as viewing AI as magic and blindly trusting AI's interpretations of their personal characteristics. Findings from this work provides a descriptive account of how people forms rationales based on their evolving AI knowledge to navigate AI misinterpretations. This study also provides design implications for responsible mitigation strategies that consider people's evolving AI knowledge to reduce potential perception harms when AI fails.

Finally, I examined students' perceptions of the AI during the ToM revision stage.

Given that AI misinterpretation is inevitable and could lead to negative student perceptions of AI (as demonstrated in my work in the ToM recognition stage in chapter 6), designing proper AI mitigation strategy is critical in enhancing students' perceptions of AI. In human-human communication breakdowns, we are able to introspect on our prior ToM process that led to the misinterpretation while repairing our mistake, and sometimes communicate about our introspection process to ease annoyance caused by the communication breakdown. Inspired by this process, I propose and examine the mitigation strategy of letting AI identify, repair, and communicate about its ToM revision by answering this question:

RQ4. How can AI's revision of its interpretation influence students' perceptions of AI in AI-mediated social interaction?

To answer this question, I devised a conceptual model of an AI agent's metacognitive module that can take in student feedback, introspect on its prior inaccurate interpretation, identify the cause of its misinterpretation, revise the misinterpretation in its knowledge base, and generate revision message that consists of step-by-step description of its revision process to provide transparency of its revision to the student. I then conducted a large-scale 3x3 vignette factorial experiment to examine the effectiveness of such revision communication of varying revision details coupled with different levels of apology sincerity by measuring students' perceptions of the AI agent. I found that mimicking human's metacognitive reasoning and communication process in revision message can make students attribute intent, emotions and other human characteristics to the AI agent. Balancing the levels of revision detail and the apology sincerity properly in the revision message is critical, given that different combinations can either elicit students' tendency to understand and forgive AI's misinterpretations, or students' eerie feelings of uncanny valley of the AI agent. This study offers concrete implications on balancing the informational and social aspects when designing AI's mitigation feedback to enhance student's perceptions of the AI.

1.4.1 Thesis Statement

My proposed theoretical framework of Mutual Theory of Mind provides a process account for understanding and designing human-AI communications that considers human perceptions of AI in AI-mediated social interaction.

1.4.2 Contribution

This dissertation contributes to the fields of human-AI interaction, computer-supported cooperative work, and responsible AI. Specifically, this dissertation makes theoretical, design, and empirical contributions by providing: (1) A set of design implications for AI systems to cater to students' social needs, alleviate students' challenges in remote social interactions, and mitigate the potential ethical concerns and risks when performing AI-mediated social interaction in online higher education; (2) The theoretical framework of Mutual Theory of Mind that posits MToM as a vision for ideal human-AI social communication where both the human and the AI can construct and respond to their interpretations of each other. This framework provides analytical power in breaking down the envisioned human-AI social communication process into analyzable stages and provides the vocabulary to describe the ToM process at each stage; (3) Design and development implications of AI systems that account for human perceptions of AI throughout human-AI communication, all based on rich empirical descriptions of the automatic construction of people's perceptions of AI, people's perceptions of AI after recognizing AI misinterpretations, and the design of AI's ToM revision reasoning and communication that can influence people's perceptions of AI.

1.5 Organization of the Dissertation

This dissertation is organized as follows. In chapter 2, I review relevant literature in Theory of Mind in human-AI interaction, human-AI communication, and AI-mediated social interaction. I then describe the study context and the AI system in my work in chapter 3. In chap-

ter 4, I describe two studies in understanding and designing for AI-mediated social interaction from online learners' perspectives through semi-structured interviews and co-design workshops. After investigating the design opportunities and challenges of AI-mediated social interaction, my work focuses on applying the MToM framework in AI-mediated social interaction in large-scale learning context to account for students' perceptions of AI throughout the human-AI communication process. Chapter 5 describes a study in exploring the automatic construction of students' perceptions of an AI agent through language analysis in the OMSCS program. Chapter 6 describes a study in understanding students' perceptions and reactions of AI after recognizing AI's misinterpretations of students' personal characteristics in AI-facilitated project team-matching in large-scale learning. Chapter 7 describes a conceptual model for an AI system to revise its interpretation of the student based on student feedback and a study that examined the design characteristics of AI's revision communication strategy to the student in large-scale learning. Finally, chapter 8 summarizes the implications of my research for designing AI's ToM-like capability to account for human perceptions of AI to achieve the vision of MToM in human-AI social communication in large-scale learning context. I discussed implications of designing human-centered AI in large-scale learning contexts, accounting for human perceptions of AI in human-AI social communication, designing the social roles of AI systems responsibly, and research opportunities in human-AI interaction through MToM beyond large-scale learning context. Table 1.1 shows an outline of this dissertation research.

Table 1.1: Outline of dissertation research.

Theme	Study	Summary	Index
Human-Centered Design of AI-mediated Social Interaction	Understanding the design space of AI-mediated social interaction in online learning [42]	Explored online learners' current practices, challenges, and needs in remote social interactions. Highlighted students' preferences for iterative communications with AI systems that can exhibit human-like social characteristics and behaviors when mediating their social interactions, which could lead to perception harms.	chapter 4
	Co-designing AI agents to support social connectedness among online students [48]		
ToM Construction: AI's Construction of Human's Interpretation of the AI	Automatic construction of students' perceptions of a virtual teaching assistant in online classes [49]	Examined the longitudinal changes of students' perceptions of the AI agent and established the feasibility for AI agents to automatically construct students' perceptions of the AI (e.g., intelligence) through linguistic characteristics (e.g., verbosity) of students' utterances to the AI agent.	chapter 5
ToM Recognition: Human's Recognition of AI's Interpretation of the Human	Examining students' reactions and perceptions of AI after recognizing AI (mis)interpretations in AI-facilitated team matching	Mixed-methods approach identified the critical role of students' evolving AI knowledge in informing their three rationales about the AI's working mechanism: AI works like a machine, a human, and/or magic. Some rationales can lead to harmful perceptions and reactions to AI misinterpretations.	chapter 6
ToM Revision: AI's Revision of Its Interpretation of the Human	Mitigating students' perceptions of AI through AI's revision of its misinterpretations	Designed and examined AI's revision of its ToM inspired by human metacognitive reasoning and communication in effectively mitigating students' perceptions of the AI. Highlighted the importance of balancing the social and informational aspects of AI's revision communication to improve students' perceptions of AI.	chapter 7

CHAPTER 2

RELATED WORK

This chapter summarizes the related work in three sections:

- **Theory of Mind in Human-AI Interaction**: I provide a brief overview of the definition, theoretical perspectives, and evaluation of ToM in cognitive science. I summarize relevant work on ToM in human-AI interaction in two subsections: building and assessing AI's ToM-like capability, and understanding people's ToM of AI in human-AI interaction. Most of such existing work treats ToM as an internal capability or internal representation instead of a process to facilitate communication.
- **Human-AI Communication**: I discuss communication process from varying disciplinary perspectives including communication studies, cognitive science, and social science. I then review existing work on human-AI communication failures and strategies to mitigate human-AI communication breakdowns.
- **AI-Mediated Social Interaction**: I define AI-mediated social interaction and review theories on designing technology-mediated remote social interaction. I highlight the importance of building social connections in online learning environment and discuss the social and ethical concerns of using AI technology in online learning.

2.1 Theory of Mind in Human-AI Interaction

2.1.1 Theory of Mind: Definition, Theory, and Evaluation

Theory of Mind (ToM) [1], our ability to infer and attribute mental states to ourselves and others through explicit and implicit verbal and behavioral cues to make predictions about behaviors, is fundamental to many human social behaviors including collaborative work

and social communications [50, 2]. ToM helps us build an understanding of people's mental states (e.g., beliefs, goals, plans, knowledge, emotions), which is critical in intentional communications, communication repairs, teaching, persuasion, and more [50, 51]. Deficiency in ToM can drastically impact people's social learning and social communication skills, as evident in people with autism spectrum disorders [2, 52, 4, 53] and psychiatric or neurological diseases (e.g., schizophrenia) [54, 55, 56, 57]. With extensive empirical and neurological backup [54, 58, 59], ToM continues to be a leading area of research in cognitive science, developmental psychology, psychiatry and other relevant fields [51, 60, 2, 1, 5, 61].

Over the years, scholars have proposed several theories to understand the internal process and definition of ToM capability. Two mechanisms have been posited to understand the process of ToM capability [62, 3]: Theory-Theory [63, 4] and Simulation-Theory [64]. According to Theory-Theory, ToM is constructed by inferring others' mental states rationally based on common sense knowledge. emphasizing on the knowledge representation and cognitive aspect of ToM [62, 3, 54, 55]. Simulation-Theory, on the other hand, posits that ToM is our innate feature or capability to "empathize" or run simulations of how others would think given certain situations and knowledge [62, 3], focusing instead on the affective aspect of ToM. While each of these two theories has gained much support, recently, some have sought to combine the two by postulating the two subcomponents of ToM: the cognitive ToM and the affective ToM [54]. The cognitive ToM emphasizes on the rational aspect of cognitive understanding of others' view, i.e., perspective-taking, whereas the affective ToM focuses on the affective aspect of sharing others' feelings, i.e., empathy [54, 64].

While the internal process of the human ToM capability remains a topic of debate, the recursive property of ToM has been studied in both the cognitive and affective aspects of ToM. Recursive ToM reasoning is the idea that I can not only infer about what you *believe* (first-order), but also infer about what you *think* about her *beliefs* (second-order). This

recursive process often presents itself in terms of "orders", depending on how many person's mental states are involved in this recursive reasoning process [65, 66]. Many ToM measurements leveraged the recursive nature of ToM to assess people's (especially children's) ToM capability, most prominent of which is called the false belief task [67, 65]. The original false belief task, proposed by Wimmer and Perner (1983), describes a story of a character Maxi, who puts chocolate into a cupboard x. In his absence, Maxi's mother then displaced the chocolate from cupboard x to cupboard y. The question to ask the subject is where would Maxi look for the chocolate when he returns? Such false belief task, including the classic Sally-Anne test as well as second-order false-belief task [65], all aim at assessing the subject's ability to have an explicit and definite representation of others' wrong belief [67], which is a critical indication of the subjects' ToM capability.

2.1.2 Building and Assessing AI's ToM-like Capability in Human-AI Interaction

The fundamental role of ToM in human-human interactions has inspired researchers to develop AI systems with ToM-like capability to facilitate human-AI interaction. Much of this effort contributes techniques and architectures to model various aspects of humans' mental states to facilitate human-AI interaction. For instance, recent work has sought to generate more transparent and explainable AI behaviors by considering *humans' mental states* [20, 68, 69]. This is accomplished by building a ToM cognitive architecture for the AI system to model possible human interpretations of AI's motions based on the interaction context [70, 71, 72, 73, 74, 37]. Such ToM module can enable the AI to generate motions that are more legible to humans [20], provide explanations that can help humans quickly understand model's strengths and weaknesses [68], as well as offer changes and explanations of the robot strategy and plans [69]. Others have focused explicitly on modeling *humans' knowledge states* during human-AI task collaborations. This could enable the AI system to account for irrational human behaviors [19], maximize human collaborator's knowledge of the environment [75], provide timely and necessary information to

humans for task completion [18], and dynamically adjust AI's own collaborative autonomy in human-AI collaborative tasks [69].

Besides modeling humans' cognitive ToM capability, others have also looked into modeling humans' affective ToM capability to enable AI systems' social communication behaviors [76, 77, 49, 78]. Modeling and attributing humans with mental states such as communication needs [76], levels of trust [77], and joint attention [78] has shown to be effective in improving the engagement, outcome, and perceptions of the AI system during human-AI interaction [76, 77, 78]. Others have pointed out the potential of leveraging linguistic cues to model people's perceptions of AI during human-AI interactions [49]. For example, researchers have inferred users' emotions towards an AI agent [79], signs of conversation breakdowns [80, 81] from communication cues. Yet whether a user's holistic perception of the AI could be constructed through linguistic characteristics extracted from conversations remains unexplored.

As researchers continue to explore various types of ToM-like capability for AI system to model humans' mental states, assessing how people perceive such AI capability becomes critical in designing human-centered AI systems. Prior work has shown that when AI systems exhibit behaviors enabled by ToM-like capabilities, such as perspective-taking [82], lie detection, or playing character-guessing games, it tends to elicit people's prosocial behaviors [82], increased acceptance towards AI [83], better engagement [84]. Such ToM-enabled AI behaviors could also lead humans to perceive the AI as more trustful [85] [86, cf.], more empathetic, and more intelligent [87, 88]. AI systems equipped with ToM-like capability can also encourage humans use of higher-order ToM reasoning in both cooperative and competitive game scenarios [89, 90, 91, 92].

2.1.3 Understanding People's ToM of AI in Human-AI Interaction

Even at the nascent of HCI when computers were not designed with human-like appearances, research has shown that people tend to attribute human characteristics and mind-

lessly apply social heuristics such as politeness to computers [93, 94]. With advanced AI systems exhibiting increasingly human-like social behaviors, people are attributing mental states such as intentions, beliefs, goals, and emotions to AI systems, or as what Dennett would call "taking the intentional stance" [95] on AI systems. Such mental state attribution behaviors are often enabled by people's ToM capability, as well as people's perceptions of the AI in human-AI interaction.

Our perceptions of the AI is a multifaceted concept that determines how we interact with the AI systems. Prior research has explored people's mental model of AI systems in various settings—in a cooperative game setting, people's mental model of AI agents could include global behavior, knowledge distribution, and local behavior [96]; people's perception of a recommendation agent consists of trust, credibility, and satisfaction [97]. When AI systems exhibit ToM-like behaviors such as profiling people's emotions [15, 98, 45] and personality characteristics [99, 47, 32, 100], people might form inaccurate expectations and perceptions of AI. Prior work found that most people perceived their AI-generated personality profiles to be "creepily accurate" [99, 47, 32, 101]. Other studies demonstrated people's tendency to over-trust AI-generated personality profiles about them. Studies showed that people felt unqualified to modify their personality profile generated by the "expert" algorithm [47], sometimes even overriding their own personal judgments about themselves due to the belief that the AI algorithm could identify their "hidden self" and had privileged information about them [45].

Scholars have developed many theories to explain people's reactions and perceptions of AI systems [102, 103, 104, 105, 46]. Theories such as Machine Heuristic, a rule of thumb that people believe machines are logical, objective, and emotionless, and hence more trustworthy than humans, has been used to explain people's tendency to over-trust AI outcomes [106, 107, 108, 109]; the Computers Are Social Actors (CASA) paradigm has been used to explain people's social reactions to forgive, tolerate, and justify AI misfires [105, 106]; learning science theories such as conceptual changes [110] and ontological

shift [111, 112] have also been used to examine how people conceptualize the ontological differences between humans and computers [46]. Together, these theories suggest people's increasingly blurred conceptualizations and reactions between machines and humans due to technologies' human-like behaviors and capabilities [46].

Other than people's perceptions of AI, research also suggested other factors that could influence people's mind attribution behaviors towards AI systems. Much work has examined children's mind attribution behaviors towards social robots and found that parental view on social robots [113], children's age [114], and children's ToM capability [115] could all influence whether children attribute minds to the AI systems. Other scholars have found that people with limited folk theories about the data source, data scope, and personalization of the AI algorithm tend to dismiss AI's personality inferences as less threatening [32].

AI's appearances and behaviors can also influence people's mind attribution behaviors to the AI systems. Prior work suggested that AI agents' humanlike physical appearances could elicit people's mind attribution behavior, leading people to perceive the AI as more humanlike, more sociable, and more amicable [116, 117, 118]. Others added that if the AI agents' physical appearances are not strongly similar to humans, then it might not able to gain attribution of mental functions given that they are perceived as different entities [119]. Besides physical appearance, AI agents' social behaviors can also influence people's mind attribution behaviors. Prior work has shown that AI agents' gaze [120, 121, 118, 122] [123, cf.], gestures [122, 120], emotions [121, 124], language cues [120] and proxemic behaviors such as physical proximity to humans [123] can elicit cognitive (e.g., intelligence, competence) and affective (e.g., emotion, empathy) mind attributions [124, 123, 121, 118, 120]. Shank *et al.*[2019], through an extensive qualitative studies of people's self-reported personal encounters with perceiving minds in AI, have summarized that people's mind perceptions are often related to their expectations of AI's abilities [120], whether AIs inhabit social roles, and AI's physical and behavioral anthropomorphic qualities [125].

People's mind attribution behaviors to AI can also influence human-AI interaction out-

comes. While children's mind perception of AI doesn't seem to affect their behaviors when interacting with social robots [126, 127], adults with more developed ToM capability could behave differently. For example, lack of mind perception of AI could enable people to exploit AI agent's perceived lack of ToM in competitive games [128]; yet perceiving mind in an AI agent could improve human performance on tasks even when the AI agent only acted as an observer [129]. Mind attribution to AI agents could also influence how people view the effectiveness of AI's repair strategies during mistakes: study has shown that apologies and denials in AI's repair strategy are more effective when people attribute more consciousness to the AI agent [130].

2.2 Human-AI Communication

2.2.1 Theoretical Perspectives of Communication

Communication is commonly defined as *"the process of transmitting information and common understanding from one person to another."* [131] Scholars across disciplines have offered different perspectives to study and enhance communication.

In communication studies, researchers have focused on the different components at play during the communication process. The classic Shannon-Weaver model of communication [132] outlines several key components during the communication process [131]: *sender* who initiates the communication process by sending messages *encoded* using symbols, gestures, words, or sentences through a chosen *channel* to the *receiver*. While the message is transmitting through the channel, there could be *noises* that could distort the message. After receiving the message from the sender, the receiver will *decode* the message into meaningful information, depending on how the receiver interprets the message. Finally, the receiver will provide *feedback* as a response to the sender. These key components determine the quality and effectiveness of the communication.

The Cognitive Science perspective of communication highlights the critical role of ToM [1]. ToM enables us to make suppositions of other's minds through verbal and be-

havioral cues, acting as the foundation of human-human communication [50, 2]. From this perspective, both interlocutors during communication can form interpretations of what's on the other interlocutor's mind based on the implicit and explicit communication cues. For example, we can often infer the interlocutors' goals, plans, or preferences based on what they said, their facial expressions, or their bodily expressions [1, 50]. Based on that interpretation we formed about the other's mind, we will act accordingly to correct, explain, or persuade. This cycle of building an interpretation of other's minds and then act upon that interpretation continues iteratively throughout the communication process. Inferring about each other's minds through behavioral cues, according to this perspective, is therefore crucial to a smooth and successful communication.

Communication process can also be interpreted from the social science perspective through impression management [30]. In his seminal work, Goffman describes social interaction as an information game between individuals and their audience to maintain the "veneer of consensus" to keep the conversation going and to avoid awkwardness. During social interactions, the audience usually try to gather as much information as they could about the individuals they interact with in order to elicit a desirable response from the individual; whereas individuals put up performances through two kinds of expressions—— expressions that are intentionally performed to leave a certain impression (expression given) or expressions that are unintentionally given off that could influence the audience's impressions of them (expression given off)—— to manage impressions [30]. Throughout interactions, each party conveys their definition of the situation through communications: individuals by expressions and audience by reactions to the individuals.

2.2.2 Communication Breakdowns in Human-AI Communication

Communication breakdowns happen frequently during human-human communication due to various factors. Lunenburg (2010) suggested six types of communication barriers that could be present based on the six components of Shannon (1948)'s communication model:

sender barrier (e.g., failure in initiating communication), encoding barrier (e.g., language barriers), medium barrier (e.g., not familiar with the communication medium), decoding barrier (e.g., lack of knowledge in popular slang), receiver barrier (e.g., not paying attention to the conversation), feedback barrier (e.g., failure to respond to a comment) [131]. While these barriers in information transmission during communications could lead to breakdowns, the social aspect of communication also plays a crucial part in a successful communication. Research has found that deficit in social cognition, such as the lack of ToM, could explain people's inability to recognize and recover from communicative failures [57]; mismatch in communication behaviors could also cause confusions and frustrations that lead to communication breakdowns in human-human communications [30].

Human-AI communication research has also offered several taxonomies on communication breakdowns and shed light on the reasons behind these communication breakdowns. Paek (2003) created a taxonomy of communication errors that spans across fields in human-human and human-AI communication by detailing four levels of coordination for grounding mutual understanding in communications: channel level (attempt to open a communication), signal level (understanding what behavior is intended as signal), intentional level (understanding semantics of the signals), and conversation level (in which a response is generated) [133, 134]. Building upon this taxonomy, Hong *et al.* (2021) created a taxonomy of natural language failures for each level: attention (channel level), perception (signal level), understanding (intention level), and response generation (conversation level). In human-robot interaction literature, Honig and Oron-Gilad (2018) categorized robot failures broadly into technical failures and social failures. Technical failures in human-AI communication often includes cases such as the AI agent unable to perform certain action or speech [135, 136, 137] or the AI system incorrectly interprets user input [22, 135, 138]. Social failures in AI systems usually include the AI system unable to interpret user's intent [138, 139, 140] or failed at setting appropriate user expectations [141, 138, 22].

2.2.3 Repairing Communication Breakdowns in Human-AI Communication

Prior research has offered different types of repair strategies that users tend to employ when attempting to repair communication breakdowns in human-AI communications. Some work focuses on repair strategies that users employ in an attempt to fix communication breakdowns with conversational agents: research found that users tend to change their utterances through hyperarticulation (e.g., speaks louder and/or slower towards the chatbot) [142, 140], simplification of language [142], message reformulation [139, 140], restarting or repetition [142, 140] or quitting [140, 139].

To mitigate the negative consequences of AI mistakes in human-AI communications, researchers have looked into various recovery strategies for AI systems to repair its relationship with the users [135, 143, 144, 145]. These AI recovery strategies can be roughly categorized into: *Confirmation* or acknowledgement of the failure, providing *information* such as explanations to elucidate the situation, integrating human-like *social* characteristics such as apology, *disclosure* of AI's limitations and capabilities, *repairing* the mistakes, *asking* for clarifications from the user, or *delegate* to human assistance [143, 144, 145]. Based on these mitigation strategies, existing work has designed and examined various strategies' effectiveness through measuring user perceptions of the AI systems after they erred [146, 147, 148, e.g.]. However, these mitigation strategies are often studied independently in human-AI communications [149] despite the demonstrated potential of repairing communication breakdowns by combining several mitigation strategies together [150, 149].

One particular mitigation strategy that has gained popularity is explainable AI. Among the eight repair strategies that [22] suggested, three of them are dedicated to help users understand the AI system's working mechanism better by highlighting the keywords that the AI system extracted or provide explanations to specific words [22]. However, empirical studies have suggested that common explanation strategies were not as effective as expected given that the explanations did not take into account people's domain expertise [151] and AI knowledge such as intuitions and beliefs [152]. This has sparked the area of human-

centered explainable AI [153, 154, 155] that aims at presenting explanations relative to people's knowledge, capabilities, and beliefs.

2.3 AI-Mediated Social Interaction

2.3.1 AI-Mediated Communication and Social Matching Systems

AI-mediated social interaction is at the intersection of two CSCW sub-fields: AI-Mediated Communication (AI-MC) [156] and social matching systems [31]. AI-MC is defined as "mediated communication between people in which a computational agent operates on behalf of a communicator by modifying, augmenting, or generating messages to accomplish communication or interpersonal goals [156]." Existing work in AI-MC has focused on AI-augmented text communication such as smart replies [157, 158, 156] and offered valuable insights into the social and ethical challenges of AI-MC. Several research has found that AI-MC in text communication can influence interpersonal dynamics such as perceived trustworthiness [159] as well as users' perceived agency and responsibility in the communication process [157]. AI-MC could even undermine the social attraction between two human communicators due to positivity bias [158]. The issue of user agency has been frequently brought up in recent literature [103, 156]. Researchers share the concern of AI-MC usurping user agency instead of augmenting it [103, 156, 160] due to the high level of proactivity that current AI-MC systems are designed [103, 160]— sometimes humans are left out of the decision-making progress completely (e.g., auto-correct). To resolve the issue with user agency, anthropomorphic AI systems such as social robots and conversational agents have been suggested to help users gain a sense of agency since the interactions with anthropomorphic AI systems heavily depend on user responses to take further actions [103]. In a recent review of AI-MC research, Hancock *et al.* also point out the social and ethical implications of AI-MC such as the potential issues of using AI systems to dictate and enforce a certain communication style, as well as concerns surrounding disclosure and transparency of AI-MC [156].

While AI-MC can be incorporated into a variety of technologies across contexts, I envision a possible scenario of integrating AI-MC into social matching systems due to the inherent social nature of communications. A social matching system is a particular type of recommender systems that aims at providing recommendations of people that might be of interest for someone to connect with [31, 161]. Social matching systems, while prominently used in the online dating context [162], have also been employed to rediscover old friends on social networks [163, 164], link job-seekers with potential employees [165] and connect academic researchers to local community collaborators [166]. Social matching systems thus offer a new way for individuals to build their social capital [161, 167], satisfy people's needs to socialize [31, 168], offer opportunities for chance encounters [169, 161, 170], and potentially reduce human biases during the matching process [165].

Social matching systems have been traditionally evaluated through accuracy and efficiency [171], however, there have been growing calls for human-in-the-loop evaluation and assessment [171, 172, 31, 165]. Terveen and McDonald urge future research in human-centered social matching systems to explore the need for transparency in systems' decision-making process as well as the balance between match accuracy and user privacy. These concerns are also echoed by other relevant literature, calling future research to explore explainability and user privacy in social matching systems from a user-centered perspective [171, 172]. While the basic functionality of social matching system is often to recommend people with commonalities, researchers have pointed out the potential ethical consequences of creating echo chambers and polarization in the community by reinforcing people's similarity-seeking behaviors [165].

With the advancement of natural language processing and recommendation algorithms, AI-mediated social interaction that combines the features of AI-MC and social matching systems could present new design challenges and opportunities. However, we currently don't have a clear understanding of user's perspective on the design and ethical implications of AI-mediated social interaction— the intimate nature of social matching combined with

the intrusiveness of AI-augmented social profile could raise more ethical concerns than either AI-MC or social matching systems.

2.3.2 Technology-Mediated Remote Social Interaction: Theories and Design

Decades of CSCW research has produced many prominent theoretical frameworks to guide the design of technologies in supporting remote interactions, which is at the core of CSCW research. Among these theoretical frameworks, Ackerman (2000)'s social-technical gap [173] and Erickson and Kellogg (2000)'s social translucence [174] both draw inspirations from in-person interactions to design technology that can support remote interactions.

Ackerman (2000) defines social-technical gap as "the great divide between what we know we must support socially and what we can support technically [173]." In his seminal work, Ackerman (2000) points out that when technology mediates remote interactions, they are often designed to be rigid, reductionist, and do not allow sufficient ambiguity compared to in-person interactions [173]. Much research has since adopted this framework to identify the social-technical gap in a variety of contexts such as health tracking [175], collaboration among telesurgery teams [176], online collaborative consumption [177], and many more.

While the notion of social-technical gap typically acts as a general guide and call-to-action for CSCW research to bridge this gap between social and technical requirements, Erickson and Kellogg (2000) go a step further and outline detailed principles on designing towards socially translucent systems to support natural online interactions [174]. Specifically, Erickson and Kellogg (2000) propose that socially translucent systems should have three characteristics: visibility, awareness, and accountability. *Visibility* refers to system's ability of making social information more visible; *Awareness* refers to people's ability to know each others' existence; *Accountability* refers to system's ability to hold people accountable for their behavior by generating and enforcing social rules. Erickson and Kellogg (2000) believe that these three characteristics allow people to observe, imitate, aware, and interact with others socially in in-person context, and thus building socially translucent sys-

tem is a fundamental requirement for people to carry out normal interactions online [174].

Much research has explored the design and implementation of social translucence in technology-mediated interactions across a variety of contexts. The most common implementations of social translucence is through building social proxy [174] and collective awareness [178]. Building social proxy to implement social translucence was first described in the original Erickson and Kellogg (2000) paper, in which they present a design of the "Babble" system that demonstrates user presence and activities through a simple graphical representation. Social proxy is later integrated into system architectures used to support Wikipedia knowledge workers [179]. Collective awareness is also a crucial design factor in socially translucent systems [178, 180]. Prior research has posited methods to support collective awareness through creating common repository to generate mutual understanding for members of globally distributed teams [181] and conducting synchronous coding sessions for learner engagement [182].

However, despite their prominent roles in guiding the design of technology-mediated interactions, to our knowledge, these two theoretical guidelines have not been empirically examined for guiding the design of AI systems that can facilitate social interaction among online learners. It remains unclear of whether and how AI-mediated social interaction could fulfill the requirements of social-technical gap and social translucence in online learning context.

2.3.3 Social Interaction in Online Learning Environment

Building strong social ties among online learners has long been recognized as a crucial factor to improve students' satisfaction [183, 184, 185], reduce dropout rates [186], and stimulate intellectual exchange by providing a safe atmosphere [187, 185]. However, online learners frequently report feeling socially isolated [35, 34, 43]. With the increasing demand of online learning in higher education, much research has offered strategies that could help improve online learners' social presence and sense of social belonging in online learning

environments [188, 15]. Most of these strategies center around what the instructors could do (e.g., share personal stories, use humor and emoticons), what the students could do (e.g., contribute to discussion boards), and how the course design should be changed (e.g., limit class size, structure collaborative learning activities) in order to foster social presence for online learners [188, 189, 190]. Despite the increasing call for research on building and designing technologies to address this issue in the online learning environment, we have seen very few technical systems that explicitly focus on helping online learners build social connections.

One of the reasons behind this lack of existing technologies to help online learners build social connections is the tendency to restrict social interactions to academic tasks that are often learning-oriented, in which social interaction is often only in service of obtaining desirable learning outcomes [191]. For example, online learners reported that working on group projects together helped them get to know other students on a more personal level and discover affinities [43]. However, these relationships are often ephemeral and don't usually last beyond group projects— once the common educational goal of completing a group assignment is gone, online learners often go their separate ways [43]. This pitfall is also reflected in the design of CSCW technologies intended for online learners, the majority of which aims at facilitating online learners' cognitive learning processes, such as tools to facilitate peer discussions [192], encourage help-seeking in online discussion forums [193], crowd-edit lecture videos [194], and facilitate student teamwork [195].

For the HCI and CSCW communities, research that aims at examining how technologies can be designed to foster social connections among online learners is only at its nascent stage. A few initial studies have tried to understand how online learners currently build social connections through extensive interviews [43] or short surveys [35]. These studies found that online learners form lightweight social connections through the discovery of shared identity [43, 196, 35], commonalities among online learners such as location in the same city, in either the self-introduction thread or through working on the same group

projects [43, 35]. Class-oriented activities such as group projects seem to be one of the few opportunities for online learners to interact closely with each other and identify affinities, with few students building relationships that last beyond the class [43, 35]. However, there is a lack of systematic and in-depth investigation into the design of technologies that could cater to online learners' existing difficulties in social interactions and help facilitate their social interaction process.

2.3.4 Social and Ethical Concerns of AI in Online Learning

AI technology is no stranger in online learning contexts— growing and prominent fields such as Learning Analytics and Educational Data Mining have successfully collected and analyzed online learners' digital footprints with the purpose of understanding and enhancing students' learning outcomes and environments [197, 198]. With all the learning activities happening on digital platforms in online learning, much of students' data are readily available— researchers have been able to collect online learners' online behavioral data such as clickstream data [199, 200, 201], educational records [202, 203], demographics [204, 205], online discussion posts [206, 207, 208], even facial expressions and physiological data [15, 201] in order to analyze and enhance students' learning process. Based on these data, researchers were able to predict online students' learning performance [198], provide decision support for teachers and learners [209, 15, 198], detect students' behavioral patterns for learner modeling [198, 202], as well as predicting and identifying online students who are about to drop out [206, 198].

While there are many benefits in using AI technology to gain insights into and advance online students' learning, ethical concerns have been raised regarding the large-scale of data collection, monitoring, and analytics on students' data through AI [210, 15, 211]. Privacy is among the top ethical concerns given most online students are probably not aware of the extent their data is being collected and analyzed [210]. Since most of the student data are automatically recorded by the online learning platforms, students have limited freedom

in controlling what data gets collected [210, 212]. Even if students were given control over sharing of their data, prior research suggests that for students, the perceived benefit of improving their learning outcomes often outweigh the cost of sacrificing their data [210]. Combining this cost-benefit analysis with students' high levels of trust in giving their data to schools [15, 211], researchers have cast doubt on commonly proposed solutions to protect students' privacy such as informed consent and terms and conditions [210].

Besides privacy concerns, scholars have also brought up issues with the interpretation and validity of learning analytics results [210]. Learning analytics researchers have admitted that we often don't have all the necessary data to provide valid interpretation of students' learning behavior [210, 213]. Another issue is the transient nature of students' identities, which renders the inferences made by AI technology frequently outdated and invalid [210]. Misinterpretation of students' learning analytics data often result in misdirected learning intervention which counters the goal of enhancing students' learning experiences [210].

I note that almost all of these concerns regarding the use of AI in online learning center around data collection and analysis with the purpose of improving students' learning. When students' data are being collected and analyzed by AI technology for social purposes, a different set of ethical and social concerns could surface. For example, prior social media research has shown that people carefully manage their social images online [214, 215] and more scrutiny is required when making social and emotional inferences based on people's online footprints [216]. Additional concerns could also be raised when the AI system deployed are AI agents that often more capable of eliciting more private and sensitive personal information (e.g., credit card information [102]) from the users during interactions [217, 102, 218].

CHAPTER 3
STUDY CONTEXT & AI SYSTEM

As higher education continues to scale up to meet adult learners' needs to upskill and reskill for career developments [38], more and more adult learners are experiencing social isolation due to the lack of face-to-face and spontaneous social encounters [43, 34]. While instructors have devoted efforts to encourage social interactions among students such as assigning group projects or organizing class discussions [191], adult learners are only able to form ephemeral connections that don't usually last beyond group projects [43]. Considering adult learners' education goals of career development and transitions while juggling additional responsibilities from their full-time jobs and families, such class-based ephemeral connections often fall short of supporting online learners' needs throughout the program, leading to high drop out rates and low satisfactions [38, 185, 184, 186]. However, large-scale learning environments also provides an opportunity for AI systems to promote social interactions based on adult learners' needs by inferring their identities and goals through the massive amount of student data scattered around the online discussion forums and class group chats. This makes large-scale learning environments a promising application context for AI-mediated social interaction.

This thesis examines MToM in AI-mediated social interaction in the context of large-scale learning contexts in higher education. While this thesis in its entirety examines students' perceptions and perspectives of AI in large-scale learning contexts including large in-person classrooms and online learning, much of the initial studies in this thesis, specifically chapter 4 and chapter 5, were conducted in the context of online for-degree education programs, specifically the Online Master of Science in Computer Science Program (OMSCS) at Georgia Tech. To gather concrete design implications for AI-mediated social interaction, I also introduced, designed, and deployed the AI agent SAMI (stands for "Social

Agent Mediated Interaction") in several studies in this thesis to elicit students' reactions and perceptions of AI systems mediating their social interactions.

In this section, I provide an overview of the OMSCS program, including the program's brief history, the student demographics and background, as well as a description of the AI agent SAMI for AI-mediated social interaction.

3.1 The OMSCS Program as An Exemplar of Large-Scale Learning Context

As of Fall 2021, the OMSCS program[1] offered through Georgia Institute of Technology had more than 11,000 students enrolled, with around 38% international students coming from 100+ countries all over the world. The program consisted of 20% women and 11.4% under represented minorities. The size of each class varies from 200-1000 students per class. The average age of a student starting at this online program was 30.

Many students enrolled in this program were working through their degree part-time while working full-time jobs (87%). Students' goal of enrolling in this program consists of career advancement (74%) and career transition (38%) with 33% seeking community of peers. Many students in the program did not have a computer science degree (around 70% of the students do not have an undergraduate degree in computer science) but have some level of programming experience.

Students in this program used a variety of communication tools, but mostly through online class discussion forums and the program's Slack channels, initially started by students themselves. To facilitate students' social interactions, the instructor of the classes often started self-introduction threads on the class discussion forum and encouraged students to introduce themselves on the discussion thread. Students typically included the following information in their self-introduction posts: locations (city, state, country), previous and current jobs, previous and current classes they have taken, hobbies, program goals.

[1]For more details about the OMSCS program see: https://omscs.gatech.edu

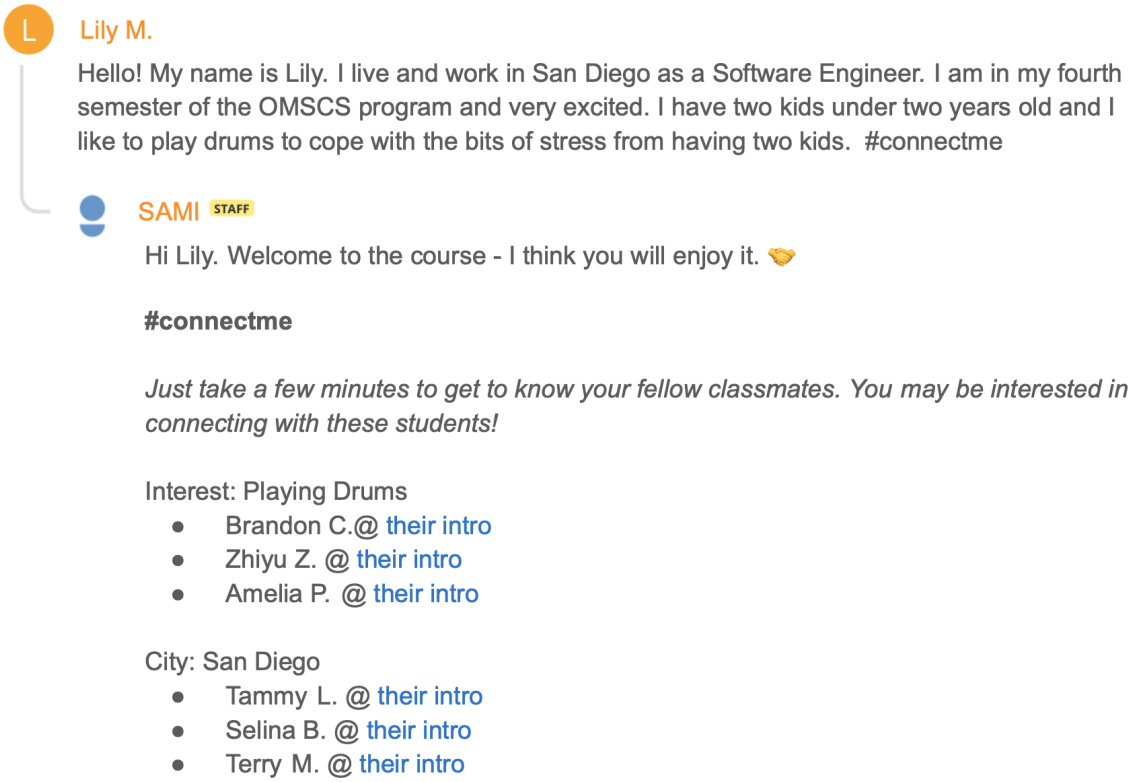


Figure 3.1: An example interaction between student and SAMI. Names in this interaction are pseudonyms.

3.2 AI Agent SAMI for AI-Mediated Social Interaction

SAMI [36] is an AI agent designed to facilitate social interactions among online learners. SAMI performs AI-mediated social interaction through extracting and analyzing the entities embedded in students' self-introduction posts posts (e.g., interests, hobbies, locations, etc.) on the discussion forum, then reaches out to the student who opted in to use SAMI and provide personalized social recommendation. Over the years, SAMI's development team has tried different methods of connecting students together by composing a social recommendation that listed students' initials, commonalities, link to the students' original self-introduction post, or by inviting students with similarities into private groups on the discussion forum. Figure 3.1 shows an example of the interaction between SAMI and the student.

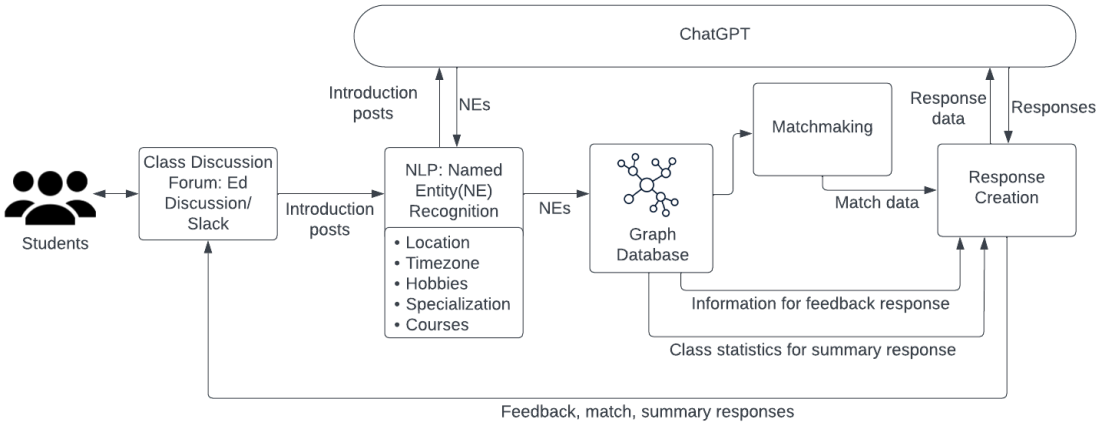


Figure 3.2: SAMI's most recent architecture with ChatGPT integration as of summer 2024. Figure taken from Kakar *et al.* (2024).

To perform AI-mediated social interaction, SAMI was designed with five modules: (1) a module that hosts the class discussion forum API; (2) a natural language processing module that uses Named Entity Recognition to extract relevant entity information from students' self-introduction posts; (3) a knowledge base module that stores all the relevant entities for each student; (4) a matchmaking module that finds social matches for the student based on their shared identity in the knowledge base; and (5) a response generation module that composes and generates the social recommendation to each student. Figure 3.2 shows the detailed and updated SAMI architecture as of Summer 2024. More detailed SAMI functionalities, architecture, evaluation, and design can be found in Kakar *et al.* (2024).

From the lens of MToM, SAMI's AI-mediated social interaction process can be interpreted as SAMI makes inferences about the students' profile information, forms interpretations of each student and stores these interpretations in the graph knowledge base, and then generate a communication feedback reflecting SAMI's interpretation of the student. Figure 3.3 shows parts of SAMI's graph knowledge base, which is SAMI's ToM representation that consists of its interpretations of the students.

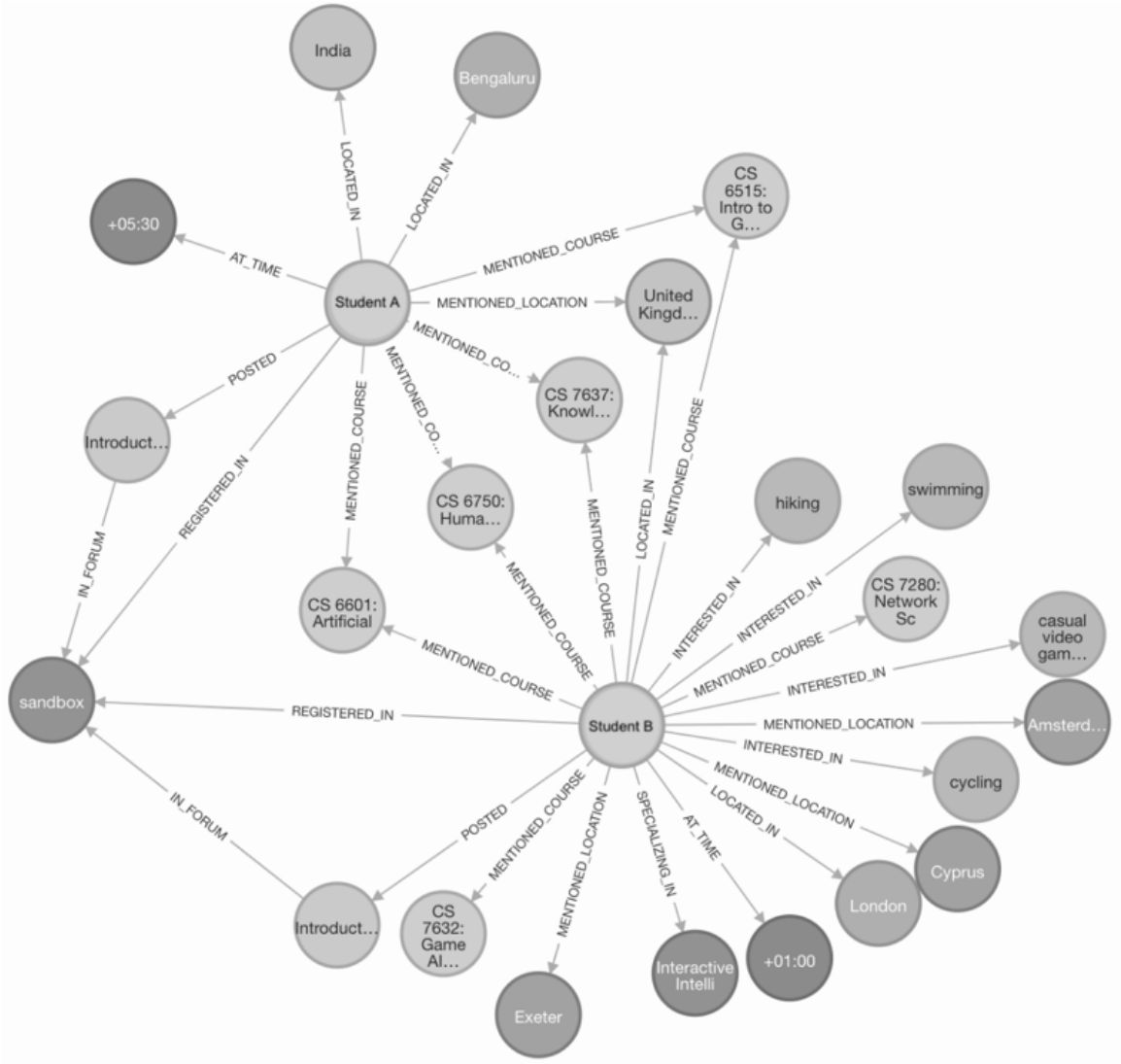


Figure 3.3: A snippet of SAMI's graph database (SAMI's interpretations of the students), taken from Kakar *et al.* (2024).

CHAPTER 4

HUMAN-CENTERED DESIGN OF AI-MEDIATED SOCIAL INTERACTION

This chapter examines the human-centered design of AI-mediated social interaction in the context of online learning as an exemplar of large-scale learning environments.

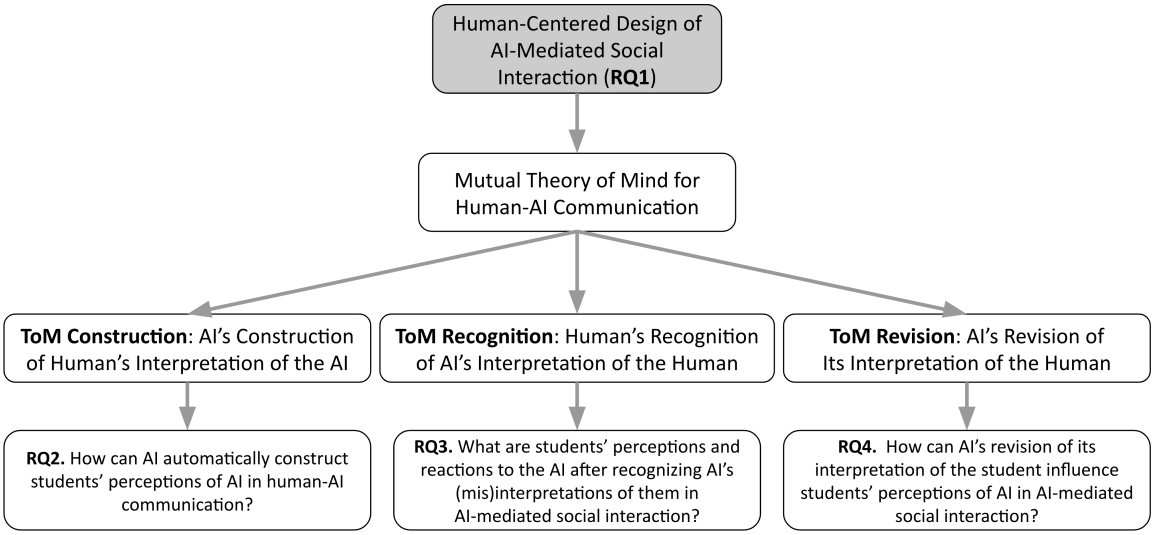


Figure 4.1: Chapter 4 investigates the human-centered design of AI-mediated social interaction.

Online education has become a common application context for AI systems to leverage student data to enhance students' learning experiences [38]. However, most of these AI systems are designed to improve teachers' teaching presence and students' cognitive presence during learning, with little dedicated to enhance the significant lack of social presence in large-scale online learning context [39, 38, 15]. Social presence, like teaching presence and cognitive presence, is a critical part of online learners' learning experiences [39]. Lack of social presence and social belongingness can lead to negative learning experiences and social isolation for learners in online learning context [39, 43]. From a human-centered perspective, I seek to design AI systems that are *useful*, *usable*, and *ethical* [219] in mediating students' social interactions to enhance online learners' social presence. To do this, it is

critical to understand online students' existing practices, difficulties, and needs in remote social interaction for the AI system to be *useful*; to empower and engage online learners in the design process to derive design implications for specific AI tools for the AI system to be *usable*; and to investigate the potential ethical and social concerns online students might have about AI systems for it to be *ethical*. Following this approach, this chapter examines this research question:

RQ1. What are the design requirements of AI-mediated social interaction from online learners' perspectives?

To answer this question, this chapter presents two qualitative studies conducted with online students at Georgia Tech's OMSCS program. In the first study (section 4.1), I conducted semi-structured interviews with online students to understand their existing practices, challenges, and needs in building remote social connections, following the deployment of SAMI at the OMSCS program. I identified design opportunities and challenges for human-like AI systems to mitigate online learners' existing social challenges and cater to their social needs through AI-mediated social interaction. Building upon findings from the first study, the second study (section 4.2) sought to pinpoint specific design implications for human-like AI systems by including online learners as active participants in a series of co-design workshops. Through these workshops, I derived specific design implications for human-like AI agents' functionalities, social characteristics, and ethical concerns when mediating online learners' social interactions. These two studies highlighted the potential for human-like AI systems such as AI agents to mitigate students' social challenges in online learning, and established the need to design AI systems that can account for students' perceptions of the AI to prevent harms in AI-mediated social interaction.

4.1 Understanding the Design Space of AI-Mediated Social Interaction in Online Learning

4.1.1 Introduction

AI-mediated social interaction is a CSCW sub-field that sits at the intersection of AI-Mediated Communication (AI-MC) [156] and social matching systems [31]. While existing work has examined the design, ethical and social challenges of AI-MC, social matching systems, and CSCW technology [165, 157, 156, 31, 173, 174], the design opportunities and challenges of AI-mediated social interaction remain largely unexplored. For instance, several classical CSCW theoretical frameworks point out the importance of designing technologies to bridge the social-technical gap [173] and bringing social translucence [174] in online interactions, yet whether and how AI-mediated social interaction could fulfill these requirements remain unknown. AI-MC and social matching literature suggest several ethical and social challenges such as privacy [31], agency [165, 156, 157, 103], trust [157, 156], and transparency [156, 31, 103]. Situated at the intersection of AI-MC and social matching, AI-mediated social interaction could raise more challenges by using AI to enhance an already personal process of building social connections. Before AI-mediated social interaction becomes more prevalent, it is critical to examine and consolidate its design space through understanding its challenges and opportunities.

With the rapid advancement of AI, AI-mediated social interaction will soon be utilized not only to help people find social partners but also to create and facilitate social cohesion within online communities. One potential application as such is the use of AI-mediated social interaction in the context of online education, where fostering social connections is not only paramount to learners' success but also urgently needed to improve learners' online learning experience [187, 220]. Strong social ties among online learners are crucial to improving students' satisfaction [184, 185], reducing dropout rates [186], and stimulating intellectual exchange [187, 185]. Conceptual frameworks of online learning such as

Community of Inquiry consider students' social presence as an integral part of successful online learning [39, 220, 221]. However, it remains unclear what difficulties online learners actually encounter during their social interactions in online learning, which makes it challenging to design AI-mediated social interaction that could cater to students' difficulties and needs.

The present work seeks to understand the design space of AI-mediated social interaction in the context of online learning. With this goal in mind, I take a human-centered approach to first understand the perceived difficulties online learners face in remote social interaction, then explore the challenges and opportunities in designing human-centered AI-mediated social interaction. Specifically, I explore three research questions:

RQ 1.1: What difficulties do online learners encounter in remote social interaction?

RQ 1.2: How can we design AI-mediated social interaction to resolve online learners' difficulties in remote social interaction?

RQ 1.3: What are the ethical and social challenges in designing AI-mediated social interaction in an online learning environment?

To examine these research questions, I conducted a qualitative study in the OMSCS program at Georgia Tech. I deployed SAMI in three online class discussion forums to help match students based on specific commonalities. I used SAMI as a probe to elicit design feedback from students based on their real experience with AI-mediated social interaction. I then conducted semi-structured interviews with 26 online students who had interacted with SAMI to understand their difficulties in remote social interactions as well as their experience and feedback on SAMI.

Through these interviews, I identify online learners' difficulties in building remote social connections, specifically, the lack of social translucence and the existing social-technical gap in current online learning platforms. Findings reveal how SAMI augmented social translucence in an online learning environment yet did not fully close the social-

technical gap. I also identify several ethical and social challenges students had about SAMI, including user privacy and agency. Building upon these findings, I discuss how to design AI-mediated social interaction to resolve online learners' difficulties in remote social interactions. Based on students' perceived concerns about SAMI, I highlight the design tension between AI performance and ethical design in AI-mediated social interaction. I then discuss the design opportunities of AI-mediated social interaction in building human-AI collaborative social matching and creating artificial serendipity to mitigate potential ethical and social challenges.

4.1.2 SAMI Versions and Functionalities in This Study

In this study, I deployed SAMI in three online class discussion forums to help match students based on specific commonalities. I used SAMI as a probe to elicit design feedback from students based on their real experience with AI-mediated social interaction. I chose to use an AI agent as the AI system to mediate social interaction among online learners due to AI agents' prior success in providing informational, emotional and social support within online communities [49, 40, 222, 223]. In this study, SAMI was designed to run on the class discussion forums, where online learners usually conduct class-related discussions and post self-introductions at the beginning of the semester. On the discussion forum, students can post questions and answers freely in an asynchronous format. Instructors can also make announcements and answer students' questions on the forum. The layout of the discussion forum is similar to typical online forums, where all the posts are organized chronologically on the left side of the screen, with pinned posts and the newest posts at the top. People with instructor access can appoint students into private groups in the form of a post thread where students in the group can communicate freely. The private groups and the posts in each group are only visible to group members. The groups appear on the left side of the screen along with all the other class discussion threads. For the purpose of this study, SAMI was granted instructor access to put students into different private social groups.

Table 4.1: The functionalities and example student-SAMI interactions of the three different versions of SAMI.

Version	Functionalities	Example Interaction
SAMI 1 (Past)	Reads and processes students' self-introduction posts on **the self-introduction thread organized by the instructor**. Students can opt-in by including #ConnectMe in their self-introduction posts. Posts and updates a separate thread that provides a **summary of class demographics** by timezone, hobby, etc. For example, "23% of the class is in CST timezone...". Generates **individual replies**.	**Student**: "Hi everyone! I'm Sarah. I live in Chicago. I am taking Computer Networks this semester. I enjoy traveling and hiking." #ConnectMe. **SAMI 1**: "Hi Sarah. Welcome to the class! #ConnectMe You may find it helpful to connect with some other students in this course. These students are also taking Computer Networks this semester: • Anthony N. - link to Anthony's self-introduction • Susan L. - link to Susan's self-introduction These students are also interested in hiking: • Mary I. - link to Mary's self-introduction • Tom S. - link to Tom's self-introduction"
SAMI 2 (Current)	Reads and processes students' introduction posts posted on the dedicated **"Introduce Yourself to SAMI" thread**. Generates **individual replies**. **Invites students to private groups** created for students with commonalities. **Posts ice-breakers** in the group to help students start conversations, as shown in Figure 4.2.	**Student**: "Hi everyone! I'm Sarah. I live in Chicago. I am taking Computer Networks this semester. I enjoy traveling and hiking." **SAMI 2**: "Hi Sarah. Nice to meet you! Fun fact: This semester, I've already met 7 other students interested in traveling, 12 other students taking computer networks. Are you interested in connecting with any of your fellow classmates? • We have identified potential connections relating to your **region, city, timezone, and hobby**. • Decide whether you like to connect with peers by replying with some, all, or none of the bolded terms above. " **Student**: "city, hobby, timezone" **SAMI 2**: "Based on your reply, I have connected you with the following groups: Chicago, timezone-CST, hiking, traveling. When you get a chance, pop into the 'Welcome' threads to say hi!"
SAMI 3 (Future)	Reads and processes **everything students posted** on the class discussion forum. Generates **individual replies to any posts** student posted that indicated their interest on certain topic and SAMI 3 was able to find students with shared interest.	**Student**: "I just found this article on robotics that looks very interesting. Thought I'd share with all of you. " **SAMI 3**: "Hello! I noticed that you are interested in robotics. Would you like me to connect you with other students who are also interested in robotics?" **Student**: "Sure! That sounds great!" **SAMI 3**: "Awesome! I just added you to the robotics group. You can access the group via this link."

In this study, onely one version of SAMI (SAMI 2) was deployed and I presented three different versions of SAMI to participants in the interview to better prompt for students' preferences of AI systems that can mediate their social interaction process. Specifically, I presented the previous version of SAMI (SAMI 1), the current version of SAMI (SAMI 2), and a future version I envisioned (SAMI 3). Table 4.1 described the functionalities and example interactions of the three versions of SAMI. While all versions of SAMI can extract useful information through the textual data on the class discussion forum to perform matches between students, there are some notable differences that I want to highlight here. SAMI 1 only provides a list of students with commonalities to each individual student while SAMI 2 and SAMI 3 directly put students into groups. SAMI 2 also posts ice-breakers in the group to help students start the conversations. SAMI 1 posts and updates a running thread that summarizes class demographics while SAMI 2 only provides a short summary that is related to individual student's profile. SAMI 1 and SAMI 2 are restricted to collecting information from students' self-introduction posts while SAMI 3 can collect and infer information from all the posts online students share on the class discussion forum.

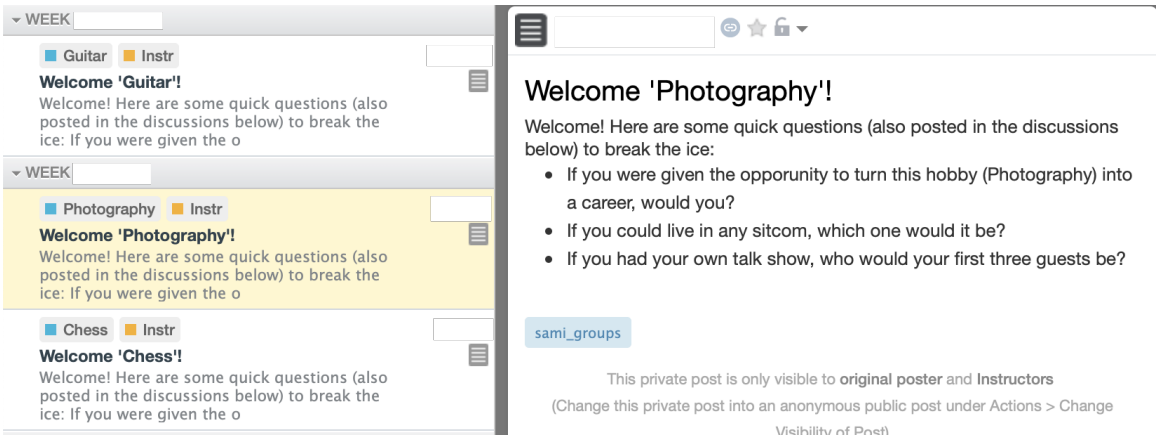


Figure 4.2: An example of the groups SAMI 2 created to help connect online students. Note that this screenshot reflects the view from SAMI's account. Students can only see the groups that they are a part of. SAMI 2 also posts ice-breaker questions in the group, as shown in the figure, to help online students start the conversation.

4.1.3 Data Collection

To understand online students' difficulties in building social connections with others and understand the design space for AI-mediated social interaction, I conducted 26 semi-structured interviews with online students enrolled in the three online classes where SAMI 2 was deployed. A detailed breakdown of the interview participants' information is shown in Table 4.2. Participants were recruited through purposeful sampling [224]. I first identified students who indicated their willingness to participate in the interview study through recruitment questions inserted in the standard class survey at the beginning of the semester. During the initial stage of recruitment, I randomly picked a batch of potential participants and sent them email invitations to participate in our study. As I interviewed more students, I then purposefully invited students of certain gender or seniority in the program to ensure I obtain perspectives of a diverse set of students.

The interviews focused on understanding online learners' difficulties in building social connections with other online students in the program (the full interview protocol can be found in Appendix A). I also discussed students' experiences with SAMI 2 to gain design implications on building AI systems that can help facilitate online learners' social interaction process. During each interview, I started out by understanding online learners' experiences in the online program, which included their goals in enrolling in the online program, their general experiences in the program, and their study routine. Next I asked about online learners' current experience in building connections with other students, specifically what kind of interactions they have had, their preferred types of interactions, and the difficulties they have encountered in building online social connections. Finally, I asked about online learners' experiences with SAMI and their feedback on the AI agent. To help participants better articulate their preferences, I presented three versions of SAMI (listed in Table 4.1) and asked about their likes, dislikes, and potential concerns for each version of SAMI.

All the interviews took place virtually after SAMI was active on the discussion forum for at least six weeks. The interviews lasted from 47 minutes to 95 minutes, with an average

length of 64 minutes. All the interviews were audio-recorded and later transcribed.

Three researchers analyzed the interview transcripts through open-coding [225]. We conducted three iterations of coding and collaboratively distilled themes emerged throughout the coding process. The first iteration of coding was conducted by the three researchers in a line-by-line fashion and resulted in 74 low-level codes. The codes in first iteration stayed close to the meaning of each sentence, for instance, “don’t want to be the first one to reach out” and "timezone differences". During the second iteration, through continuous discussions, we ended up with 20 codes such as “lack of visibility” and ”desire SAMI to be more human-like”. In the final iteration of coding, we ended up with 10 categories that highlight online learners’ current and desired types of interactions, challenges encountered when building connections with others, and design implications for building conversational agent to mediate their connection process. Throughout the entire coding process, the three researchers met and discussed the codes on a regular basis and resolved conflicts that arose in each iteration of the data analysis process.

4.1.4 Findings

Through my analysis, I found that deep social connections were rare yet highly desired among online learners. However, there were a number of difficulties that online students encountered when trying to establish social connections with each other. I present these difficulties through the lenses of social translucence and social-technical gap. The interviews also revealed that SAMI was able to augment social translucence to some level yet not able to fully bridge the social-technical gap in online social interactions. We also identified a set of challenges and concerns that students had about SAMI in AI-mediated social interaction. I discuss the findings in detail below.

Table 4.2: Interview participant information. “M“ stands for “Male”, “F” stands for “Female”. “Country” column indicates the countries that participants were born in. The “# of Classes Completed” column indicates student’s seniority in the program. Online students in the program usually take 1 or 2 classes per semester.

ID	Gender	Age	Country	# of Classes Completed
P1	M	18 to 24	United States	0
P2	F	25 to 34	United States	2
P3	M	45 to 54	India	7
P4	M	35 to 44	Uzbekistan	7
P5	F	25 to 34	India	0
P6	M	25 to 34	United States	3
P7	M	25 to 34	United States	7
P8	M	35 to 44	United States	1
P9	M	25 to 34	South Africa	0
P10	M	25 to 34	United States	0
P11	M	25 to 34	United States	0
P12	M	25 to 34	China	4
P13	M	25 to 34	United States	3
P14	M	25 to 34	United States	6
P15	F	18 to 24	United States	5
P16	M	25 to 34	United States	7
P17	F	18 to 24	United States	3
P18	M	25 to 34	Canada	3
P19	F	35 to 44	China	2
P20	F	35 to 44	Cuba	3
P21	F	25 to 34	China	1
P22	M	35 to 44	Iraq	3
P23	F	25 to 34	Canada	4
P24	F	18 to 24	Lithuania	5
P25	F	25 to 34	India	2
P26	M	35 to 44	Argentina	7

Perceived Difficulties in Remote Social Interactions Among Online Learners

The interviews revealed that online learners had little social interactions with each other, especially private social interactions. About half of the participants reported having only interacted with other students academically (P3, P5-14, P18, P21, P23), through either working on group projects together or discussions on public discussion forum threads. The other half of the participants managed to keep in touch with one or two other students in the program and checked in with each other on a less than frequent basis (usually monthly). These private connections were usually established through the discovery of shared identities or experiences during previous group projects (P1, P16, P17, P19, P22, P26), meetup events organized by the program or students themselves (P4, P15, P24), class communication channels (P20, P25, P26), or common experiences (e.g., working at the same company, attended the same undergraduate institute) outside of the program (P1, P2, P15).

While not all private connections developed into deep friendships, establishing close interpersonal relationships like friendships was highly desired by majority of the participants in our interviews (P1, P2 P5-7, P9, P10, P12, P14-P18, P20-26) to reduce the feelings of social isolation and to offer emotional support and social support. Yet only eight participants reported building friendships through the program (P2, P4, P16, P19-P21, P24, P25). Through the interviews, I further investigated the difficulties online learners encountered while attempting to build social connections with each other. I found that these perceived difficulties were largely due to the lack of social translucence offered by the online learning platforms— the reduced visibility of social information, the diminished awareness of potential social companion, and the decreased accountability in social behaviors. Another set of difficulties stemmed from the existing social-technical gap in online social interactions— the lack of randomness and spontaneity that is inherent in in-person social interactions but difficult to replicate in online learning environments. I describe these difficulties through the lens of social translucence and social-technical gap below.

Reduced Visibility of Social Information. Reaching out and building connections with strangers can be an intimidating process. During in-person interactions, we are able to gain social cues from other people's behaviors, gestures, or facial expressions, however, all of these social cues become invisible in an online environment [174]. This is exactly what happened to online learners when they tried to establish social connections with others.

I found that the reduced visibility of social information, such as others' willingness to connect, often left students hesitant in reaching out or maintaining the connections with others. Many participants mentioned that not knowing whether other students were willing to talk or build connections with them made it difficult to initiate conversations with others (P1-P6, P12-P14, P17, P19, P20, P22, P23). P23 said, *"When I go introduce myself on the introduction thread on the forum, I can also see other people's introductions. But then it's also a little bit vague on the signal on whether they really want to connect with someone."* While online learners were able to meet others through class group projects or the discussion forum, students said they were not sure whether others wanted to maintain the connection with them. P14 mentioned that he really bonded with some of the students in his previous group projects, yet when asked about maintaining the connection after the group project was over, he said *"When I thought about reaching out to some of the folks I've worked with before, my first thought was like 'Oh no, you'd be imposing', 'That was just a group they wouldn't want you reaching out...' That's just my first reaction whenever I thought about it."*

Communicating with other students on public channels such as Slack and the class discussion forum was one of the main methods to get to know other students and to make students themselves known to others in the program. Yet for some students who described their own personalities as more reserved (P1, P3, P5, P7, P11, P13-P15, P17, P19, P20, P22, P23, P25), the reduced visibility of how their messages could be perceived by others added pressure when they wanted to reach out to other students. This pressure sometimes even prevented them from posting messages on public communication channels. P23 said

"I know some study groups exist but then I don't really have a desire to join them to some extent. I think it might be because I'm doing it online. Sometimes you feel intimidated to join a big group and then posting your view on certain things since you don't know anyone in the group." For female students who were minorities in the program, the pressure was even higher when posting on public channels. P15 said *"Being a minority in this program, sometimes it makes me feel even more nervous about posting anything because I don't want to represent females badly. I don't want to post something bad or stupid on the class discussion forum and other people would be like 'Oh that's one of the very few women in this program. The woman in this program must be dumb.' I don't want to be a bad representative. I don't necessarily feel intimidated to ask questions or to talk to people because I'm a female student, but it can make it harder to relate to people."*

Diminished Awareness of Potential Social Companion. One of the main goals and advantages of online learning is to help education scale up by giving more students the opportunities to learn [226]. The downside of this is that online classes usually have hundreds or even thousands of students per class. This has resulted in online students' diminished awareness of other students' existence in the program when building social connections. In my study, participants talked about how the diminished awareness of potential social companions posed challenges in their social connection process, specifically, the difficulty of identifying potential social companions and the lack of personal touch in the online learning environment.

The overwhelming number of students and activities within each class made it extremely difficult for online students to identify others that they could potentially build social connections with (P2, P3, P6-P9, P11, P12, P14-P16, P19-P21, P23-P26). P14 said, *"The Slack group and the class discussion forum can get overwhelming depending on what's going on with all the posts. Looking for specific students to connect with is like trying to find that needle in a haystack."* P21 also said, *"There are many many posts in the public forum. I may never find the person or the group I'm interested in if I search it manually. "*

With hundreds of students communicating via the class discussion forum and Slack group, these main communication channels could quickly become walls of text, which made all the interactions there seem impersonal. The lack of awareness of human characteristics in a majority of the communications provide a weak foundation and decreased motivation for students to build social connections. When asked about the students' self-introduction thread for each class, P2 said *"The introductions students gave were really nice. But those discussion threads get so overwhelmed and the content of the introductions are people's names, how many classes they have taken, current classes they are taking, where do they work. So it just becomes a wall of text repeating 'Oh I live in New York', 'I live in San Diego'. That's not super valuable to build a strong relationship. It is just nice to see the reminder that 'Right, there are people here'."* P15 also pointed out the missing personal aspects that were inherent in in-person interactions *"The program is huge and it's so hard to differentiate people unless you meet them in-person. That's part of the reason why I like meeting people in-person. There's not as much of space to talk to people about things other than classes. Even though we all relate to each other since we are in the same class, but it's harder to get to the actual personal aspects where you relate to each other."*

Decreased Accountability in Social Behaviors. Erickson and Kellogg point out that while awareness and accountability often co-occur in physical world, they are not usually coupled in the online spaces. Accountability is often fostered through the creation of social norms in the online environment that hold people accountable for their social behaviors [174]. In my study, I found that both the existence and the lack of social rules could prohibit online learners' social interaction processes in online learning environments.

Some implicit social rules could restrict or deter students' social behaviors in online spaces. In my study, I found that online learners designated different purposes for the communication platform that they commonly used— the class discussion forum was for academic discussions and the Slack group was for casual interactions (P1, P3, P4, P8, P10, P11, P18-P22). While the intention of the class discussion forum was to replicate the phys-

ical classroom where students could have interactions and discussions about and beyond academics that could facilitate student learning as well as building social connections, the implicit social rules of only using the online class discussion forum for academic discussions seemed to restrict students' social interactions with each other on the forum— students felt accountable to only have academic discussions on the forum instead of casual conversations. Considering most of students' interactions tend to happen on the discussion forum, this social rule significantly limited online learners' opportunities for building social connections. For example, P19 said, *"I prefer Slack if it's just casual conversation. I don't feel casual conversations are okay on the class discussion forum. The forum is for more serious conversation for the class. So I'm not going to post any unrelated things on the forum."* Other students also felt like the class discussion forum was monitored due to the presence of the teaching staff. P22 said, *"The thing about the class discussion forum is that it's not friendly. You don't feel open to post on the forum or maybe that's me. At least I feel like the discussion forum is official and monitored. If I said something wrong on the forum people would judge me for it. "*

While working in group projects with other online learners provided some social pressure for students to interact with each other in small groups, the pressure was gone once the project was finished. Even though the same thing could also happen in in-person classrooms, online students didn't usually run into each other again after the class was over. The lack of repetitive encounter with each other in the online program thus reduced students' feeling of accountability to talk with each other again (P1, P3, P6-P8, P11-P15, P19, P21-P23). P8 said, *"Last semester we had a group project with five students. We had our own Slack channel to communicate and at the end of the semester we were all very friendly. It would be nice to work with them in the future but there's no more intersection of us. Even if we ended up in the same class, there is no way for me to know that because I can't look through everyone's name in my class. I think most likely we are not ever going to be in the same class again so there is no place for us to interact again. "*

The Social-Technical Gap in Remote Social Interactions. Besides the diminished visibility, awareness, and accountability in building social connections in online learning environments, another set of difficulties that online learners faced was the lack of spontaneity and randomness in online environments. While social interactions are inherently spontaneous and random in in-person interactions, participants reported that in online environments, they had to intentionally make efforts to compensate for this social-technical gap [173] in remote social interactions.

Many participants (P1, P3, P5-P7, P9, P10, P13, P14-P16, P21, P26) reported that online interactions were not as spontaneous and organic as in-person interactions. P15 pointed out the importance of having "in-between" moments during interactions, which proved to be difficult to achieve in online environments: *"Sometimes when you meet people, you have those in-between moments, where you are not necessarily actively working on the project, but you are still thinking about the project. I really valued those moments. So I really wanted to be able to meet up with my teammates in person and have that joy in getting to know someone and then have those in-between moments."* Other participants also pointed out that during in-person interactions, work conversations often organically led to more social activities. P6 said, *"My current interactions with other online students are more academic or professional. It wasn't like my friends in undergrad. I think that's one of the other things that's odd about the social interactions with online program. In undergrad I can make friends, go have dinner, we can go out and get a drink or whatever. That kind of very natural social interaction can happen, which I don't see the analogy online."*

On top of the lack of spontaneity in online social interaction process, several participants (P1, P5, P8, P10) talked about the random encounters that on-campus students could have that offered starting points for them to build social connections. However, these random encounters were almost completely missing from the online program. P8 described different scenarios where random encounters could happen in in-person campus, *"Let's say you are on an actual college campus and you go to the library to study. You might end*

up being in a situation where you can talk to someone who is in the same university but in a different class or major... Or the university has some open lawn that sits between the lecture halls and the food court. So people would walk through there everyday and that's a place where you can run into someone. So to translate that into online program, I think it's hard to generate a place that students have to go to."

The lack of spontaneity and randomness of online social interactions forced students to spend extra effort and time to build those social connections (P1, P3-P9, P11-P15, P17, P20, P23, P26). P5 said, *"So when I was in college, I never went out to form connections like 'Oh let me form five new connections'. It just organically happened in the process of studying."* Many students also had to go out of their way to form those connections, such as driving for 40 minutes to meet up with other online students in the same city, emailing every student in the class to build connections, or manually looking up students with commonalities among thousands of discussion forum posts.

The general consensus among participants was that in online learning programs, the social and learning aspects were often separated compared to the traditional in-person educational programs. Instead of forming social connections organically during the process of taking classes or walking around campus that were inherently built into the on-campus educational experience, online learners had to establish social connections in a more intentional way.

Augmenting Social Translucence in Online Learning through SAMI

Based on students' experiences in interacting with SAMI, I asked for participants' feedback on SAMI. Through my evaluation of SAMI, I found that SAMI was able to augment social translucence in online learning environment mainly by improving the visibility of social signals and increasing students' feeling of accountability in remote social interactions. While SAMI raised students' awareness of potential social companions to some degree, participants pointed out several ways on how SAMI could further improve their awareness

throughout the process.

Enhanced Visibility of Social Signals. I found that SAMI made social signals more visible among online learners, especially in reference to highlighting students' willingness to build social connections (P1, P3, P5, P23). Specifically, participants highlighted the feature in SAMI 1 and SAMI 2 that made it easier for them to infer students' willingness to connect. For SAMI 1, the #ConnectMe tag was intended to allow students to opt-in for SAMI 1 to connect them with other students in the class introduction threads. However, online students interpreted this as a signal of whether students were willing to build social connections. P23 commented on the #ConnectMe function in SAMI 1, *"I like SAMI 1 a lot, especially the #ConnectMe. This is sending the student a signal that there's people who are interested in chatting with other students. So I think this is very helpful."*

Students were also able to infer students' willingness to connect from SAMI 2. For example, P1 said, *"People who introduced themselves to SAMI 2 are more likely to want to connect to other people. So it's a group of people that are likely to be more willing to talk to other people."* P23 also said, *"There are Slack groups that can form similar groups that SAMI 2 did for us so forming a group is not a problem here I feel. Knowing who is available to form a group or engaging people who are interested to do certain things is a challenge. SAMI 2 made this process automatic so this is great."*

Improved Awareness of Potential Social Companions. In online classes with hundreds of students per class, SAMI also raised awareness of potential social companions for online learners by highlighting students' shared identities (P1-P5, P7, P9-P19, P22, P23, P26). P21 commented on the feature of connecting students based on similarities in all three versions of SAMI, *"SAMI's useful because manually, I may never find the person or never find the group I'm interested in. But SAMI can find some related groups or students I might want to connect with for me. "* Participants also said even without SAMI's feature of directly connecting students together, the class demographic summary statistics posted by SAMI 1 was

also useful in offering the lost personal aspect to online learning, *"I really appreciate SAMI 1's class demographics summary. I think it's interesting to see how many people are from different places and in different classes. It's cool to see just how broad the classes are and where everyone is from."*

Some students also suggested that the next generation of SAMI should connect students together based on more specific commonalities, instead of the current broad shared identities (P2-P4, P9, P11, P15, P17, P23, P26). For example, P2 commented on SAMI connected her with other students in the San Francisco area, *"That's great but it's not something you can make a strong connection with. Bay Area is huge so even if there is someone else in the Bay Area, they could live literally two hours away from me. So you really have to narrow in the location."* P4 also suggested SAMI to connect students based on multiple shared identities instead of only one, *"Let's say I want to meet people who are also interested in hiking, I wouldn't want to connect with people who are located outside of my city. Because sure we could probably connect on the forum and share some past experiences, but that would probably be it. Hiking is not something you discuss online, it is something you do outside of online environment."*

While SAMI gave students awareness of potential social connections, many students pointed out the necessity of continuing to support that awareness throughout the entire process (P3, P6, P7, P12, P16, P17, P20, P22, P24). Participants expressed their preference to know more about the students recommended in SAMI's reply, instead of just how many students share similar commonalities. For example, P17 commented on SAMI 1, *"I like how SAMI 1 is sort of pointing out, like here's the list of people you might want to connect with. I like how it links to the students' posts in the introduction threads. I think that's useful for quickly seeing 'Oh this is what this person said about themselves'."*. SAMI 2 did not provide information about each student in each SAMI group, which diminished students' awareness of others in the group. P22 suggested, *"Maybe if there is a student join the group, SAMI will say 'hey everyone, we have a new member just joined the group. This*

is John Doe. Please say hi to him. John Doe please introduce yourself."'

Greater Accountability in Social Interactions. SAMI also provided some level of accountability by offering some structure to the social interaction process (P1, P3, P5, P6, P14, P15, P16, P19, P21, P22, P24). By putting students directly in groups, SAMI provided the push for students to start interacting with each other and helped alleviate students' mental barriers on having to reach out to other people. For example, P3 compared SAMI 1 and SAMI 2, *"I think SAMI 2 is a better approach because it actually places you in the group as opposed to SAMI 1. I would probably not going to initiate any communication with the list of people SAMI 1 recommended to me because you typically don't initiate communication with somebody directly for no purpose right? So SAMI 2 creates that sense that you are already in a group and that you can share something more at ease."* P15 also said, *"I like SAMI 1's reply, but I do think it puts a lot of pressure on the students to have to reach out. I feel like it gives you a lot of information but it might not be the push that people need in order to reach out to others. Whereas with SAMI 2, because it makes the group so it at least starts to move in the direction of eliminating barriers that people would have to reach out to other people."*

After putting students in individual groups, SAMI 2 also posted ice-breaker questions that were relevant to the group topic for everyone to start the conversations. P24 believed this provided some kind of accountability for people to start talking since the questions were straightforward, *"I think if SAMI 2 did not post anything, it would be like, what are we supposed to do here? But having the questions that specifically relate to the topic to the group, like Seattle group 'what would you do in Seattle?' is great. Because you obviously all have that shared interest. So being able to talk about it in a structured way is very smart. "*

Challenges and Concerns about SAMI in Mediating Social Interaction

While SAMI efficiently augmented social translucence during online learners' social interaction process, we also identified several challenges and concerns that students' expressed about SAMI in terms of AI-mediated social interactions. Based on my interviews with the online learners, i found that SAMI was not able to fully bridge the social-technical gap in online social interactions and even exacerbated the gap due to its lack of human-like characteristics and unnatural behaviors. Students also pointed out that SAMI did not offer enough transparency into its working mechanism and decision-making process. When asked about concerns about the potential continuous usage of SAMI in online learning environment, students pointed out that privacy was not a perceived concern, however, there were concerns about SAMI 3 potentially excluding certain students out of the social connection process as well as students losing their agency in building social connections.

Social-Technical Gap Remains: Lack of Human-likeness and Naturalness. One of the difficulties that online learners experienced during social interactions was the lack of spontaneity and randomness compared to in-person interactions. This existing social-technical gap forced online students to intentionally spend more effort, time, and energy to build social connections with each other. SAMI helped reduce this gap slightly by automating the social interaction process and thereby reduced some efforts that students spent in building social connections (P1, P4-P7, P10, P11, P14, P17, P21, P26). P17 believed that by providing social recommendations, SAMI played the similar role to that of a mutual friend. P14 also said that since SAMI recommended social matches automatically, he didn't have the need to intentionally put himself out there to build social connections.

However, throughout my interviews with online students, I noticed that SAMI also contributed to the social-technical gap in online social interactions due to its lack of human-likeness and naturalness (P2, P4, P5-P7, P15, P18, P21-P24). While many technologies are often designed to be artificial, rigid, and reductionist, when introducing an AI agent to

operate in human communities to build social connections among community members, the agent was expected to act more human-like and natural. In my interviews, online learners believed that SAMI did not act natural enough. For example, P2 said *"I think SAMI 2 needs to feel more human-like and not as robotic to make people more comfortable interacting with it. SAMI 2's response just seems so numerical with those statistics and normal human don't really talk like that."* P7 also said, *"It would be good if it seems like it's actually interacting with me instead of an automated response. Even though it's an AI I don't want to feel like it is. If SAMI 2 acts more like a human, it would seem like a person responding to me who knows the people that are recommended. So I want it to seem like it is a person."* When asked about why they wanted SAMI to be more natural, P4 said that when SAMI acted unnaturally, it would "break" the natural atmosphere of online forum environment and therefore broadened the social-technical gap in online environment:

> *"So with the class discussion forum is that, when you have a conversation with someone, it sometimes feels like you're in the same room talking to people. Like when answers to the posts come in real time and you have this sort of atmosphere of speaking in a group. Like you are sitting in the room and talking to people and you could see if others are trying to answer your question. Maybe they do some research on the side and then they post their responses or maybe they know it right off the top of their head. So sometimes it feels like a real time conversation. So any sort of artificial-looking posts would interrupt this flow. Like if you imagine a group of students in a study group sitting at the same table studying and then there's like some TA announcer speaking some robotic automated message every once in a while, it would interrupt this flow, right? So, this sort of a robotic type of response kind of interrupts this flow on the discussion forum." (P4)*

Insufficient Transparency in SAMI's Working Mechanism. On top of SAMI's lack

of naturalness, students also reported that SAMI was not transparent enough about its decision-making process and working mechanism (P3, P5-P10, P13, P14, P16-P18, P21, P23-P25). I found that students wanted to learn more about SAMI's capability for them to better communicate with SAMI using similar vocabularies. For example, when asked about what he wished he had known before using SAMI 2, P7 said, *"I wish I had known that there was a possibility of not getting matched due to my word choices. If I were told to use these key words or some word bank to have better matching results, that would be helpful. Like if I knew to use 'travel' instead of 'traveling'."* Other students took guesses as to why SAMI 2 didn't work as they expected, such as not putting them into certain groups when it was clearly mentioned. P3 put "travel" among several other hobbies but was only assigned to "travel" group, he said, *"I'm just trying to see why I got put in 'travel' but not other hobby groups. That's why I figured maybe SAMI 2 only picked out the first hobby and place people in those groups."* P16 also had similar doubts, *"I guess I am curious as to whether I didn't get assigned to 'Virginia' group. Maybe because I said 'Nothern Virginia', so it got stuck in 'Nothern Virginia' and it thought it was different from 'Virginia'. I guess if SAMI 2 told me what it is thinking that would be great. Because SAMI 2 also didn't pick up on the courses I listed, which might be because I abbreviated the course names."* To improve transparency, P24 suggested, *"I would have appreciated if SAMI 2 had another response after my reply, saying something like 'Sure I will look for other people who are also in Los Angeles.' or 'I will let you know if I find someone else that played Ultimate Frisbee.' Just something that tells me 'I've stored this information and I'm looking for that connection."'*

Concerns about Excluding Some Students Out. While all participants were generally very positive about SAMI, some students brought up concerns that they had about the continuing use of SAMI in online learning contexts. Since SAMI only operated on one of the many platforms online learners used to engage with each other, there were concerns about SAMI inadvertently excluding some students out of the AI-mediated social interaction process among online students (P11, P12, P15, P19, P23, P26). SAMI 1 and SAMI

2 currently only took information from one specific introduction thread on the class discussion forum and SAMI 3 would also be restricted to the class discussion forum. This design thus naturally left out students who were not active on the class discussion forum and students who did not post their self-introductions. P15 commented on this concern regarding SAMI 3, *"I do think it's kind of hard to have something based off of people posting about themselves on the forum. I think the forum can be sort of self-selecting for students who are most willing to put themselves out there on the forum type of people."*

Concerns about Losing Agency in Building Social Connections Online. While students appreciated SAMI automating some of the most difficult processes in building social connections with other online learners, students also raised concerns about losing agency in making social connections online (P7, P9, P10, P11, P14, P15, P20-P22). SAMI 2's feature of putting students directly into groups was popular among the interview participants for creating accountability to interact with each other as well as alleviating the mental barrier of reaching out to other students. However, this feature, while created adequate amount of social pressure for students to initiate interactions, was also commented by the participants that it took away their agency in choosing which groups they could join. For example, P9 said, *"I would actually prefer to participate in a class community just so that I can hang around the edges of it first before committing. Because there are some things that I'm interested in doing but I haven't done it yet. "* This concern came up more after we described our vision for SAMI 3 that could more naturally and automatically connect students together based on everything they posted on the discussion forum. After hearing our vision for SAMI 3, P14 said, *"It would be great if students could opt-in to particular topics. Maybe SAMI 3 could provide its observation first then let student decide. Just some kind of mediating steps that let you sign off, like 'Oh no I would prefer not to have other people know that I'm interested in this."'* P10 also brought up the point of whether to trust SAMI 3 on the connections recommended, *"So if we were to compare this to real life, generally, when you meet someone new, you have a bit of a gatekeeper in your friend who introduces*

you guys. So you have a mutual person whom you both trust. So then because you trust that person, then you trust that the person they introduce you to is going to be not a murderer or something like that. So, yeah, SAMI 3 may not give you that."

Privacy was Not a Perceived Concern. However, even though privacy is often a concern for AI systems that leverage public data [227], many participants did not have privacy concerns regarding SAMI (P3, P5, P6, P10, P12, P14, P16, 17, P19-22, P24). Some participants believed that the intention of posting on public forum was for others to see it. For example, P12 said, *"If they already posted their self-introductions that means he/she wants that information to be public. If I don't want others to know where I work, I won't put that information there."* In fact, some students indicated their willingness to post more frequently on the discussion forum if SAMI could offer more accurate matches. After hearing about our vision for SAMI 3, P14 said, *"I think that's a really brilliant idea! I think I'd be more inclined to post too, because I don't need to worry that my posts would be falling on deaf ears. I am more willing to put myself out there more since I will be giving SAMI 3 a chance to find more things and possibly make more connections in the program."*

4.1.5 Discussion

These findings offer insights into the challenges and opportunities of AI-mediated social interaction in an online learning context. Specifically, I pinpoint online learners' perceived difficulties in building social interactions due to the lack of social translucence and the existence of the social-technical gap in their current online learning environment. The findings reveal the potential and challenges of using AI-mediated social interaction to help facilitate online learners' social interaction process.

Designing AI-Mediated Social Interaction for Online Learners

Through my in-depth investigation of online learners' social interaction process, I identified online learners' pressing needs to build close friendships with other students, yet

these needs were largely unfulfilled. The findings echo prior work in that academic-related interactions are the main sources for students to build connections with each other [43, 35]. While the discovery of shared identities or common experiences could help foster closer connections and may occasionally lead to long-term relationships [43, 196, 35], close friendships are uncommon among online learners [43, 196] yet highly desired by students seeking emotional support. Many online learners stepped out of their comfort zones to pursue an online degree that is very different from their academic background often with the goal of career shift or career advancement, which makes the online education process even more difficult for them. Therefore, social and emotional support become crucial to help online learners persist through the online program.

I also identified two sets of difficulties that online students encountered during remote social interactions: the lack of social translucence [174] and the existence of the social-technical gap [173] in current online learning platforms. Specifically, current online learning platforms do not offer enough visibility of social information (e.g., others' willingness to connect), provide limited support of students' awareness of potential social connections due to the large number of students, and do not facilitate sufficient accountability that is often necessary to create social pressure for students to connect during in-person interactions. Designing online learning platforms to be more socially translucent could mitigate a large set of difficulties that online students tend to experience during remote social interactions. Some design strategies to achieve socially translucent online learning platforms could be adding icons on students' avatars to indicate their willingness to connect with others, thus improving visibility of social cues [180]; highlighting students' shared identities through social matching features to offer awareness of potential social companions; and providing dedicated socializing sections or instant video chatting features on the discussion forums [34] to improve students' accountability in building social connections.

The second set of difficulties online learners experienced stemmed from the existence of the social-technical gap [173] in the online learning environment, which created a separa-

tion between social interactions and education activities in online learning. In in-person educational programs, students often build social interactions or long-term friendships spontaneously through random encounters that happen in between learning activities (e.g., repeatedly taking the same classes, randomly running into each other in the dining hall, spontaneously grabbing coffee together after group project meetings), this kind of randomness and spontaneity is difficult to replicate in the online learning environment— in online environments, all virtual meetings are intentional and every message has a specific purpose. Without randomness and spontaneity embedded into their online education experience, online learners are forced to spend extra effort and time (e.g., driving for an hour to meet others in person or emailing everyone in the class to connect) to intentionally form social connections. This speaks to the lack of nuance, flexibility, and ambiguity in building social interactions on online learning platforms. My work thus points out the direction for future research to explore how to integrate more social activities into online learners' learning activities to artificially create more random and spontaneous encounters in online learning environments.

The Design Tension in AI-Mediated Social Interaction

This work identifies several challenges and concerns about AI-mediated social interaction such as the possibility of excluding certain students from the the social matching process [228], the lack of transparency in AI-mediated social matching processes [229, 156], and the possibility of strengthening people's similarity-seeking behavior [165]. While some of the challenges align with prior findings in CSCW, AI-MC, and social matching space, AI-mediated social interaction, situated at the intersection of several fields, also presents another challenge that combines concerns from these fields and creates *a unique design tension between AI performance and ethical design.* In my study, this design tension in AI-mediated social interaction is manifested in two areas: *the tension between matching accuracy and user privacy, and the tension between the effectiveness of social introduction*

outcome and user agency.

This study highlights the tension between achieving high social matching accuracy and maintaining user privacy in AI-mediated social interaction. The basis of getting accurate social matches is the deep understanding of user preferences, goals, and needs, which would require the AI system to collect as much user data as possible [165, 31, 229]. In the case of AI-mediated social interaction, such information can not only be gathered through information that users voluntarily offered to the AI system, but also through analyzing prior public postings to make inferences that users don't explicitly consent to sharing [230, 231]. This renders the protective measures suggested by prior literature in social matching system such as restricting data source to public information or obtaining user consent before data collection [31] less applicable. What concerns me more is the fact that no privacy concerns were raised in my interviews with the online CS students even after I showed them SAMI 3, which would have access to all of students' prior forum postings and have the ability to make inferences about their preferences and goals— in fact, participants in my study indicated that they were willing to post more frequently on the discussion forum to improve matching accuracy. Does this suggest that designers of AI-mediated social interaction should always weigh accuracy over privacy? I think not due to the intangibility of people's understanding of privacy harms and people's tendency to trade privacy for potential gains [232]. In fact, I see this as an opportunity to further investigate what types of latent behavioral data would improve social matching results yet raise little privacy concerns for users in different contexts.

Another tension presented in AI-mediated social interaction is the balance between maintaining user agency and ensuring successful social introduction processes. The introduction process between two matches is crucial for an effective social matching process [31]. In my study, SAMI 2 directly put matches into a group on the discussion forum instead of waiting for students to reach out to others. Students found this to be helpful as several of them pointed out that they would have never reached out to their matches to build

connections otherwise. However, during my interviews, students also expressed concerns about losing their agency in choosing whom they could connect with, which is inherently a very personal decision-making process. While prior research has suggested that given the convenience and efficiency AI systems facilitate in our daily lives, some level of human agency could be sacrificed as a trade-off [103], in my study I found that students were unwilling to cede control of the decision-making process in choosing whom they should connect with. This differs from prior research in AI-MC and recommender systems where issues regarding user agency have been repeatedly brought up yet could mostly be resolved by giving users as much agency during the decision-making process as possible [103, 229, 156, 157]. Based on the findings, I note that in AI-mediated social interaction, designers should strike a delicate balance between putting enough social pressure on the matches to initiate interactions and maintaining users' agency in choosing whom they could start a conversation with.

Design Opportunities for AI-Mediated Social Interactions

Building upon the design implications and ethical challenges this study has identified, I highlight two potential design opportunities for AI-mediated social interaction: designing towards human-AI collaborative social matching and creating artificial serendipity.

There are several ethical and social challenges that we outlined in the previous section that could be properly mitigated through a human-AI collaborative social matching approach. In a human-AI collaborative decision-making process, AI systems could constantly negotiate with users on important social matching steps to take next while also creating enough social pressure to prompt users to reach out to their matches. Certain way of maintaining user agency could be achieved through the use of AI agents that display proper anthropomorphic qualities [103]. Human-AI collaboration in social matching process could also resolve the design tension of maintaining user privacy and ensuring matching accuracy. This could be done by building explainability and transparency into the AI system to con-

stantly explain what data was collected and how the data was used [233]. In human-AI collaborative decision-making processes, users' willingness to collaborate is crucial for a desirable collaborative experience and outcome [234]. Luckily, aligned with prior literature [235, 236], participants in my study indicated their willingness to understand the AI agent's vocabulary beforehand to adjust their choice of words during communication in order to improve the accuracy of matching results.

Another design opportunity for AI-mediated social interaction is to artificially create serendipity, i.e., unexpected yet meaningful encounters, to artificially create randomness and spontaneity in online social interactions. Serendipitous encounters in in-person interactions such as elevator chat or water cooler conversations often result in fruitful interactions, new collaborations, new ideas, and may sometimes lead to meaningful long-term relationships [170]. In this study, one set of difficulties that online learners encountered during remote social interactions is the social-technical gap that made the online social interaction process more intentional and less natural. Designing for artificial serendipity in remote social interactions could thus re-introduce randomness and spontaneity into the online environment by helping individuals discover unexpected meaningful relationships and potentially mitigating people's similarity-seeking behaviors in social interactions [165]. Prior research has suggested that serendipity can be created by identifying individual's preferences and social networks through mobile phone sensors [237], social network information [238], and sensing technologies in the workplace [239]. Following this direction that I highlighted based on this study, future research could explore what kinds of information could be leveraged to artificially create serendipity without raising privacy concerns.

4.1.6 Limitations and Future Research

While this work has important implications for the design of AI-mediated social interaction, it also has some limitations. First, I used an AI agent with an anthropomorphic form to mediate online social interaction in this study, but I acknowledge that AI-mediated social

interaction could be performed by other types of AI systems that do not take an anthropomorphic form. Prior research also suggests that people could act more "courteous" towards AI agents due to their human-like characteristics and thus other non-anthropomorphic AI systems could raise other types of concerns than currently investigated in this chapter. I encourage future research to replicate this study using other kinds of AI systems. Second, the online class discussion forum was heavily used in the OMSCS program that this study took place in and was also most frequently brought up by students during the interviews. However, I acknowledge that this could largely be attributed to each online program's preferences and thus other forms of communication tools such as Learning Management Systems could also be leveraged to facilitate online learners' interactions with each other in other online programs. Future research should explore alternative mediums and tools that online learners in other programs used and how they could be designed to support online learners' social interactions. Third, the interview participants in this study were all studying at the computer science for-degree graduate program. Due to their major of study in computer science, these learners might be more open to the use of technology-facilitated interactions (e.g., they did not have many privacy concerns regarding the use of SAMI). Future research should try to replicate the current study at online learning programs in a different subject area to investigate concerns and preferences of AI-mediated social interaction from learners who are less tech-savvy and/or more skeptical of technology. The findings might also not be applicable to other forms of online learning environment such as Massive Online Open Classes (MOOCs) or online learning programs at the high school or undergraduate levels. Future work should examine the social interaction experiences across different contexts to contribute more knowledge in the design of AI-mediated social interaction in online learning contexts.

4.2 Co-Designing AI Agents to Support AI-Mediated Social Interaction in Online Learning

4.2.1 Introduction

In the previous study, I outlined the design space of AI-mediated social interaction through investigating online learners current practices, needs, and difficulties in remote social interaction, as well as their experience using SAMI. The findings from the previous study pinpointed the design challenges and opportunities of AI-mediated social interaction. While there are a wide variety of AI systems out there that could mediate remote social interaction, findings from the previous study presented promising potential of leveraging human-like AI systems such as AI agents to mediate online learners' social interaction process. In this section, I describe a co-design study where I focused on examining design implications for such human-like AI systems, AI agents.

AI agents have been widely employed in online learning context to sieve through large amounts of information to provide personalized learning resources [240, 241], class logistic information [40], and social information [35, 42] to individual students 24/7. Comparing to recommender systems that often operate on digital platforms [242] (e.g., shopping recommendations on Amazon), AI agents possess human-like features such as personality and natural language generation that can potentially bring spontaneity and randomness to online interactions [173, 243, 42]. In fact, an increasing number of applications are already using AI agents to help members of a community connect— Slack applications such as *Donut* and *GreetBot* all use AI agents to promote social interactions within the Slack groups. AI agents are also widely used to deliver encouragement, care, and positive energy to people experiencing loneliness and social isolation [244, 245, 16]. However, the design requirements of AI agents that can support online learners' social connectedness is underexplored.

There are two crucial design factors to consider when designing AI agents in a specific context: their *functionalities* and their *social characteristics*. Functionality refers to what

the AI agents can do to provide support. For example, AI agents have been designed with functionalities like offering well-being advice [245], facilitating therapy [246], or mediating social interactions [42] to provide support. Social characteristic refers to characteristics that determine the agent's social skills [247]. For example, AI agents can be empathetic [248, 249] or human-like [250] when interacting with individuals. Recently, in order to provide timely and personalized support, advanced AI techniques are being integrated into AI agents to detect users' emotional and mental states through emotion recognition on user dialogues and their social media [251, 252]. Understanding the desired functionalities and social characteristics of the AI agent will help determine the specific design requirements of AI agents that can help online learners feel more socially connected.

Before inserting an AI agent into an online learning environment that could have irretrievable and immeasurable impact on online learners' lives, we have to consider the potential social and ethical concerns. There have been a plethora of AI technologies designed, developed, and implemented by leveraging the readily available data generated by online learners. Many such technologies, claiming to be developed with the goal of advancing online learners' learning experiences and outcomes, often harvest and monitor large amounts of student data, ranging from demographic information to students' social media posts, *yet have hardly engaged online learners in any of the technology design or development process.* Given the unequal power relationships [210] between online learners and institutions, online learners have little say, or even awareness, of the large-scale collection and analysis of their data [210, 212].

With the goal of designing user-centered and socially responsible AI technologies that could help promote online learners' social connectedness, I take the approach of co-design [253, 254] to include online learners as active participants from the beginning of the design process. Co-design has been frequently adopted in prior literature to understand the design of AI agents across various contexts [255, 256, 257]. Through two co-design workshop studies with 23 online learners, I provide the necessary design techniques and tools for online

learners to voice their preferences and concerns freely to envision a future where AI agents could help support their social connectedness. Through these co-design workshops, I seek to explore three research questions:

RQ 1.4: What functionalities should an AI agent possess to help online learners feel socially connected?

RQ 1.5: What social characteristics should an AI agent have to help online learners feel socially connected?

RQ 1.6: What are the potential social and ethical challenges of agent-mediated social interaction in online learning?

In this section, I present and discuss design implications and ethical challenges of AI agents that could help promote social connectedness among online learners. I describe two virtual co-design workshops consisting of four design activities used to understand online learners' preferences of AI agents' functionalities and social characteristics. I also briefly present an AI agent mockup designed to elicit perceived social and ethical challenges of agent-mediated social interaction. Based on online learners' design preferences and concerns, I establish AI agents' roles as facilitators to scaffold online learners' social interaction process and discuss social and ethical implications of agent-mediated social interaction in online learning.

4.2.2 Study Overview

To explore the design requirements and potential challenges of using AI agents to help online learners feel socially connected, I conducted two virtual co-design workshop studies using the visual collaboration tool *MURAL* and the virtual meeting platform *BlueJeans*. I adopted the virtual workshop format due to the geographical spread of online learners all over the world as well as the ongoing COVID-19 global pandemic during data collection.

Each co-design workshop study consisted of three workshop sessions with three dif-

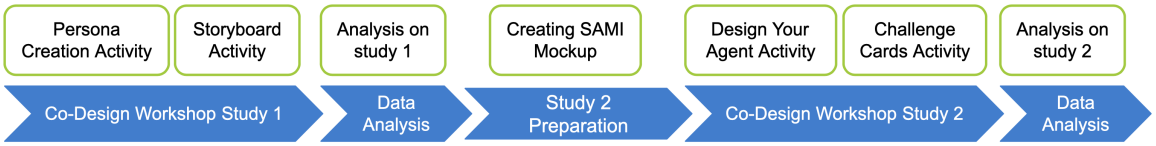


Figure 4.3: Study flow diagram that shows the different stages of our study and the components of each stage.

ferent sets of participants. I first conducted co-design workshop study 1 to understand students' desired functionalities (RQ1.4) of the AI agent through two design activities: persona creation and storyboard. I then analyzed the data collected from study 1 to identify desired functionalities of the AI agent, which allowed me to create an AI agent mockup that showcased one possible desired functionalities of the AI agent to help online learners feel socially connected. The AI agent was named SAMI [1] and the mockup was created in a storyboard format for easier comprehension of SAMI's functionalities. I then conducted study 2 to understand the desired social characteristics (RQ1.5) and the potential social and ethical challenges of agent-mediated social interaction in online learning (RQ1.6). In study 2, I explored the desired social characteristics of the AI agent through the design activity "Design Your Agent." I then introduced the SAMI mockup and used it as a probe to elicit students' perceived social and ethical challenges of using AI agents to facilitate online learners' social interaction process through the second design activity "Challenge Cards". Figure Figure 4.3 shows the overall flow of our study.

4.2.3 Co-Design Workshop Study 1: Desired Agent Functionalities

To understand the desired functionalities of an AI agent that can help online learners feel socially connected (RQ1.4), I conducted the virtual co-design workshop study 1 with three different sets of participants. The participant information of study 1 can be found in Table 4.3.

[1]Note that in this study, the AI agent in the mockup only shares a common name with the SAMI agent used in the rest of the dissertation work but does not refer to the SAMI agent.

Table 4.3: Co-design workshop study 1 participant information. "M" stands for "Male", "F" stands for "Female". The "# of Classes Completed" column indicates student's seniority in the program. Online students in the program usually take 1 to 2 classes per semester. The storyboard activity is a team activity and thus the "Team" column reflects the team composition at each study 1 session for the storyboard activity.

Study 1 Sessions	Team	ID	Gender	Age	Country (Born)	# of Classes Completed
Session 1	T1	P1	M	24	India	2
		P2	F	25	United States	1
		P3	F	26	Poland	1
Session 2	T2	P4	F	24	South Korea	5
		P5	F	25	United States	1
		P6	M	28	United States	1
Session 3	T3	P7	M	29	India	7
		P9	F	27	United States	2
		P10	M	29	United States	4
	T4	P8	M	31	England	4
		P11	F	27	United States	6

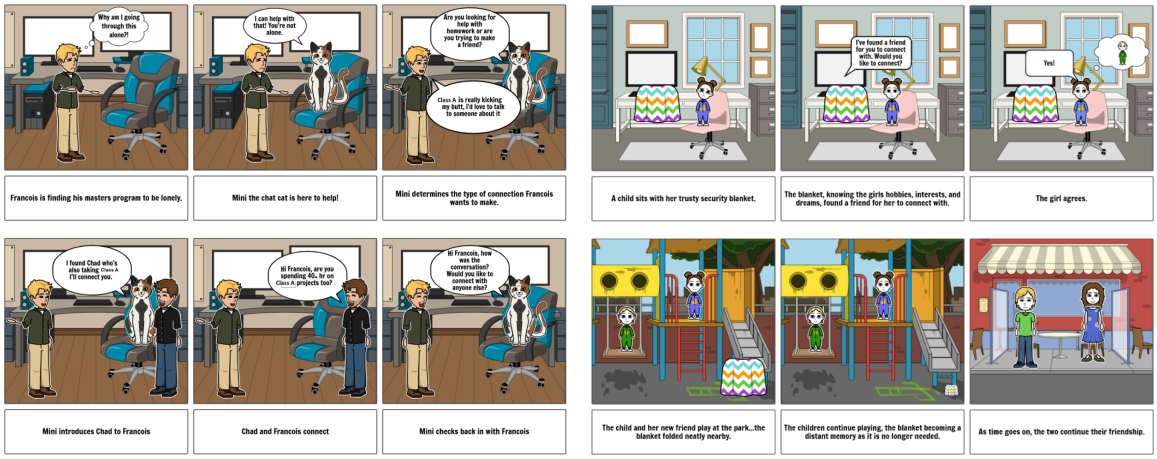
Study 1 Procedure

I began each study 1 workshop session with an introduction of the goal of the co-design workshop— to gain design implication of an AI agent that can help online learners feel more socially connected. I then introduced the agenda of the workshop session: self-introduction and ice-breaker, followed by **two design activities persona creation and storyboarding**, and concluded with a debriefing and further discussion. The worksheet that I used for study 1 can be found here.

The first design activity **persona creation** aimed at helping online learners to communicate and share their current social experiences with us and other workshop participants. I first introduced participants to the concept of persona and offered some examples of persona taken from the web that focused on different design questions (e.g., example persona for designing website to help travelers plan for their business trips). After students were familiar with the concept of persona, I provided a persona template and asked each partici-

pant to work on a persona of an online student, detailing the student's basic information, as well as his or her goals, frustration, and motivation in social interactions in online learning program. Participants were encouraged to draw on their own experience as online learners as well as other online learners' experiences that they knew of. Participants then presented their persona to the rest of the group.

The second design activity is **storyboarding**. The goal of storyboarding was to give students the method and tools to map out their desired version of an AI agent that could help online students feel socially connected. Similar to the persona creation activity, I first introduced the concept of storyboarding and provided several storyboard examples taken from existing publications on teens' creation of social robot [258] and other creative ideas of robot design [259] to demonstrate the wide range of storyboards of different sophistication, creativity, details, and functionalities. I then divided the participants into teams and put them into breakout rooms on the virtual meeting platform. The team composition can be found in Table 4.3. To help participants navigate through the storyboarding activity, I first asked each team to create a story outline following the prompt questions that were given. The prompt questions were created by us to help the team think through the interaction process with an emphasis on the AI agent's functionalities. The prompt questions were in the following order: "What makes the agent talk to you?", "How does the agent talk to you", "Where does this conversation happen?", "How does this conversation make you feel?", "Are there any actions that you or the agent need to perform outside of this conversation?", "When is the conversation over?", "What do you do after the conversation?", "How do you feel after the interaction?", "What makes the agent talk to you again?" After each team mapped out a story outline, they then proceeded to create their own storyboard on a storyboarding tool called StoryboardThat.com. I chose this tool due to its wide selection of pre-drawn scenario settings and characters, as well as its flexibility of adjusting characters facial expressions and postures. I asked each team to pay attention to these details and pick a character to represent the AI agent in their storyboard (e.g., agent could be represented



(a) Storyboard created by T3. (b) Storyboard created by T4.

Figure 4.4: Two examples of the storyboard created by the co-design workshop participants.

as an animal, a stickie figure, or many other available options) for me to get a better understanding of their expectation of the agent. Each team then presented their storyboards to the others at the workshop for further comments and discussions. I showed two examples of the storybaords T3 and T4 created in Figure 4.4 and discussed them further in the study 1 findings section.

Study 1 Data Analysis

The data I collected and analyzed included the video recordings of all three co-design workshop sessions from study 1 as well as all the artifacts created by the participants during each workshop session in study 1 (e.g., the personas, the storyboard outlines, and the storyboards).

I took an iterative and inductive approach to look at the study 1 data in two rounds of analysis. In the first round, two researchers divided up the three workshop sessions to independently review data from each workshop session. During the independent review, each researcher wrote down detailed notes on what happened during each workshop session, some interesting observations, as well as important and interesting points students made during each session. After researchers finished reviewing their own assigned sessions, the

two researchers came together and used affinity diagram to map out these detailed insights and distilled patterns and themes on a virtual collaboration whiteboard tool called *Miro*. Affinity diagramming is a bottom-up approach to organize qualitative findings in an iterative and inductive fashion. Affinity diagramming is commonly used to analyze qualitative data generated from co-design studies in prior research [260, 261]. After several iterations and discussions, we ended up with four themes and eight categories. In the second round of analysis, the two researchers switched the workshop sessions to review all data from each workshop sessions in similar fashion in the first round of analysis. Then researchers came together again and continued to group and organize new insights into the categories and themes distilled from the first round of analysis. After some reiteration and reorganization, we ended up with six categories and three themes.

Study 1 Findings: Desired AI Agent Functionalities

Online Learners' Social Connection Goals and Frustrations To explore how AI agents could help online learners feel more socially connected, I first examined online learners' social connection goals and frustrations through the persona artifacts participants created. I identified two social connection goals that were commonly shared among the workshop participants: to build long-lasting connections beyond classes (P2, P6-P9, P11) and to build connections with like-minded students who share similar interests or in similar situations (P1-P3, P5, P6). Specifically, participants revealed their desire to make friends with other online learners (P6, P7, P9), to build their professional network (P4-P6, P8-P10), and to find people who could share their online learning experiences and struggles (P7, P9-P11). However, online learners encountered a number of obstacles attempting to achieve these social connection goals. Online students found it difficult and awkward to reach out to other students that they didn't know of (P5, P7, P9-P11). In the online learning program that often had hundreds and even thousands of students per class, online learners found it difficult and time-consuming to identify students that they wanted to connect with (P1-P4,

P7, P11).

In-Situ Agent-Mediated Social Support through Continuous Monitoring. To help online learners feel socially connected, participants suggested that the AI agent should "know when" to connect the student with others— when the students really need to talk to someone. For example, in the storyboards participants created (Figure 4.4 shows two example storyboards), the agent would reach out to the student and try to connect him/her with others when the agent noticed that the student was feeling down (T1) or when the student's discussion forum posting activity was lower then usual (T2).

In order for the AI agent to conduct in-situ agent-mediated social support, participants admitted that the AI agent would need to continuously infer and monitor online students' online activities within the program. Participants suggested a couple ways to monitor student activity that could help the AI agent identify the right time to reach out to the student. For example, participants mentioned that the AI agent could monitor online learner's learning module to look for signs of frustration (T1), monitor the discussion forum activity (T2), monitor when the student log on to learning modules (T3), or even keep track of other online students' availability so that the agent could connect them with students in need (T4).

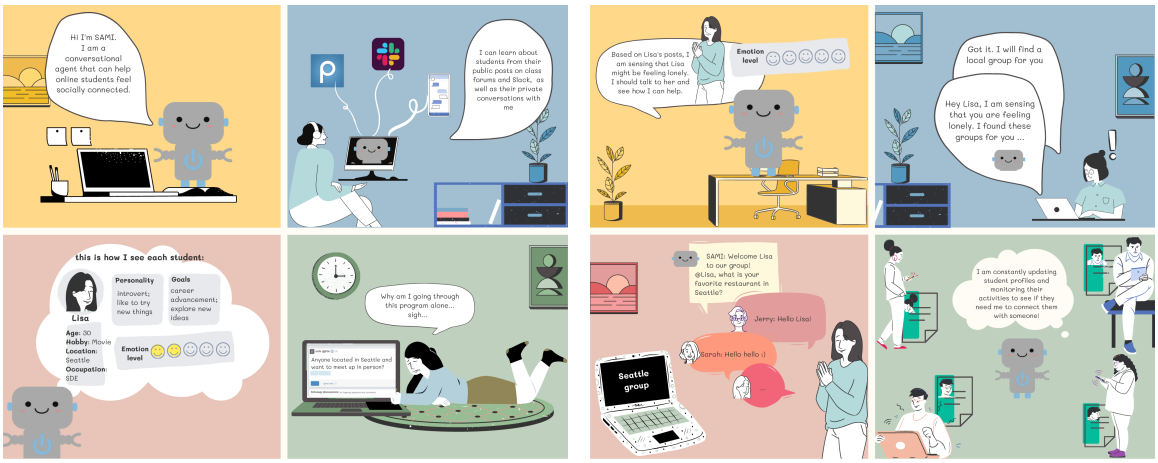
Another factor that stood out to us was the long-term and continuous nature of the agent-mediated social support demonstrated by the participants in their storyboards. All the AI agents in the storyboards interacted with the student on a continuous basis instead of a one-time interaction. For instance, both T2 and T3 said that they would want their agents to reach out to the students at the beginning of every semester or every new class. T1 and T3's agents also would check back with the students on their interactions with other students introduced by the agents.

Scaffolding Remote Social Interactions. Based on the storyboards participants created (Figure 4.4 shows two example storyboards created by T3 and T4.), I found that the AI agents were often used to scaffold online learners' remote social interaction process. As I

pointed out in the prior section that online learners encounter several obstacles during remote social interaction: the difficulty to identify like-minded students and the awkwardness to reach out to students that they don't know. To overcome these obstacles, online learners often used AI agents in their storyboards to scaffold their social interaction process through identifying and introducing online learners together based on certain criteria. For example, the agents in T1, T3, and T4's storyboards all presented the same functionalities: identified other online learners that the student would want to build social connections with (e.g., taking the same class), introduced both students together, then disappear from the conversations to let the students communicate amongst themselves. This scaffolding functionality was explicitly called out in T4's storyboard (Figure 4.4), in which the AI agent was represented as a security blanket, a comfort and transitional object that a young child often hold on to. In T4's storyboard, the AI agent, represented as a security blanket, connected the student to other students, and while the students got more familiar with each other, a.k.a., growing up together and growing out of the security blanket, the CA became a distant memory.

4.2.4 Designing the AI Agent Mockup

Based on the study 1 findings, I concluded that the agent should help online students identify other like-minded students to connect, offer continuous support to online students' changing social needs, initiate interaction with online students at the right time, and scaffold online students' social interaction process throughout. Based on these desired AI agent functionalities, I built an AI agent mockup, shown in Figure 4.5, that could help online learners feel more socially connected. In this mockup, I showed an AI agent named SAMI that could mediate the social interaction process among online learners. In our mockup, I demonstrated SAMI's ability to constantly monitor online learners' online activities within the program to understand each student's demographic information, personality, goals, their emotion levels, etc.. SAMI could facilitate online learners' social interaction by initiating



(a) Part 1. (b) Part 2.

Figure 4.5: SAMI mockup in storyboard format.

conversation with the student whenever SAMI felt like the students were feeling lonely or isolated. SAMI could reach out to the student and asked about his/her preferences in getting social matches. SAMI would then introduce the student to a group of like-minded online learners. Throughout the interaction process between SAMI and the student, SAMI was able to understand students' comments and questions and respond with natural language accordingly. SAMI continuously learned about each student's preferences and needs and reached out to them when needed.

4.2.5 Co-Design Workshop Study 2: Desired Agent Social Characteristics and Ethical Concerns

We then conducted the co-design workshop study 2, which consisted of three sessions with three different sets of participants (as shown in Table Table 4.4). The goal of the co-design workshop study 2 was to further explore the desired AI agent social characteristics (RQ2) and understand the perceived social and ethical concerns of using AI agents to improve online learners' social connectedness (RQ3). I explored these two research questions through two co-design activities: "Design Your Agent" and "Challenge Cards."

Table 4.4: Co-design workshop study 2 participant information. ”M” stands for ”Male”, ”F” stands for ”Female”. The “# of Classes Completed” column indicates student’s seniority in the program. Online students in the program usually take 1 to 2 classes per semester. The “Challenge Cards” activity is a team activity that consists of challenge teams and solution teams. The “Team” column reflects the team composition at each study 2 session for the “Challenge Cards” activity.

Study 2 Sessions	**Team**	**ID**	**Gender**	**Age**	**Country (Born)**	**# of Classes Completed**
Session 1	T5 (Solution)	P12	F	25	United States	3
		P13	M	23	United States	8
		P6	M	28	United States	1
	T6 (Challenge)	P14	M	24	United States	2
		P15	F	26	United States	3
Session 2	T7 (Solution)	P16	M	40	Republic of Panama	8
		P19	M	31	India	2
	T8 (Challenge)	P17	F	23	United States	1
		P18	F	35	United States	2
Session 3	T9 (Solution)	P20	M	28	Russia	2
		P22	M	52	United States	3
		P23	F	24	United States	1
	T10 (Challenge)	P21	F	26	Bosnia-Herzegovina	2
		P10	M	29	United States	4

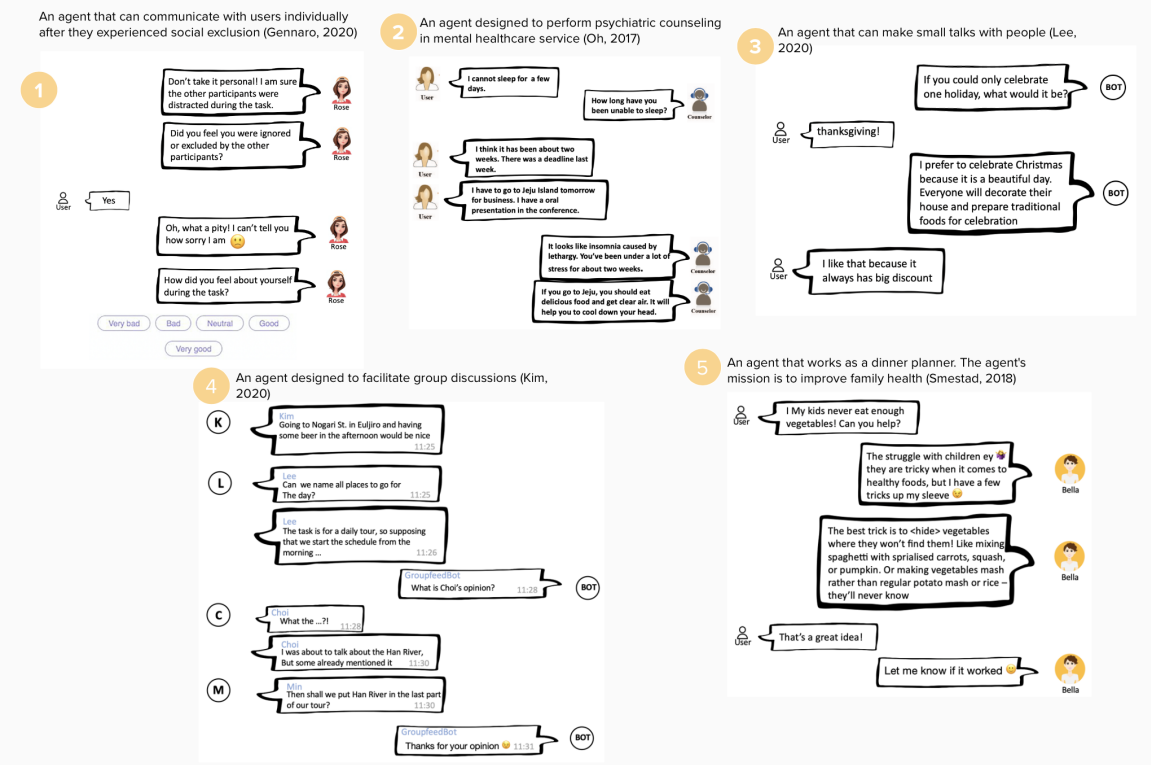


Figure 4.6: The five AI agent dialogues that were taken and adapted from prior literature [248, 251, 262, 263, 264] used in our co-design activity "Design Your Agent" in study 2. The agents in the dialogues were referred in the paper by the numbering on the upper left corner of each dialogue, e.g., "agent number 1."

Study 2 Procedure

Similar to study 1, I began each study 2 sessions by introducing the goal of the co-design workshop sessions— to understand the design requirements as well as the potential social and ethical challenges of an AI agent that can promote social connectedness among online learners. The agenda of each study 2 session includes self-introduction and a short ice breaker activity, followed by **two design activities: "Design Your Agent" and "Challenge Cards"**, and ended with a short debriefing and discussion. The worksheet I used for co-design workshop study 2 can be found here.

The goal of the first design activity **"Design Your Agent"** was to understand online learners' preferences about the social characteristics of AI agents. In this activity, I

presented to the workshop participants with five different AI agent dialogues taken and adapted from existing literatures [248, 251, 262, 263, 264]. These dialogues are shown in Figure 4.6. I chose these agent dialogues due to the fact that these AI agents possessed a variety of *functionalities* (e.g., facilitating group discussions [263], chatting with users who experienced social exclusion [248], making small talks [262])) and *social characteristics* (e.g., avatar [248, 251, 263, 264], use of emojis [248, 263, 264]) that are closely related to our vision of an AI agent that could promote social connectedness. I asked the workshop participants to discuss together and write down what they liked and disliked about each AI agent's characteristics through reading the dialogues. After participants finished discussion on all five agent dialogues, I asked participants to draw on their preferences of these agents and wrote down characteristics or features that the AI agent designed to help them feel more socially connected should definitely have or not have.

The second design activity is called **"Challenge Cards"** [265] where participants brainstormed the potential challenges and the corresponding solutions to one possible version of the AI agent that could help them feel more socially connected. In this activity, I used the SAMI mockup as a probe to elicit participants' reactions and thoughts. I first showed the SAMI mockup storyboard to the participants. I emphasized that the examples in the SAMI mockup storyboard only demonstrated SAMI's basic functionalities and that during the activity students could add on to SAMI's existing functionalities and social characteristics that they desire. I then divided the participants into two teams of two or three, the challenge team and the solution team (team composition can be found in Table Table 4.4). This design activity began with a 15-minute brainstorming session for each team in the separate virtual breakout rooms: the challenge team brainstormed all potential concerns and challenges that SAMI might elicit and the solution team brainstormed all the benefits and desired characteristics that SAMI had. After the separate brainstorming session, participants all came back together to the main virtual meeting room and began several rounds of competitions. For each round of competition, the challenge team first posted a challenge

card consisted of one potential concerns or challenge of SAMI from their brainstorming session. The solution team then tried to come up with a solution to that challenge card, drawing upon the desired characteristics and features of SAMI that they came up with in their brainstorming sessions. At the end of each round of competition, the workshop facilitator came up with follow-up questions to dive deeper into the reasoning behind those challenges and solutions that participants came up with.

Study 2 Data Analysis

The data I collected and analyzed included the video recordings of all three co-design workshop sessions from study 2 as well as all the artifacts created by the participants during each work shop session in study 2 (e.g., likes and dislikes of the agent dialogues, preferred social characteristics of the AI agent to promote social connectedness, brainstorming notes from the challenge team and the solution team, challenge cards and solution cards during the challenge cards competition).

Our data analysis process for study 2 data is the same as the study 1 data analysis process. I conducted two rounds of data analysis. In the first round, two researchers divided up the co-design session materials and conducted independent review, then came together and used affinity diagram to map out insights and distilled patterns. At the end of the first round, we had three categories and eight themes. In the second round, the two researchers swapped the workshop sessions for further independent review, and then came together to group and organize new insights. We ended up with two categories and six themes.

Study 2 Findings: Desired Agent Social Characteristics

Anthropomorphism. A crucial characteristic for AI agents in general is their level of anthropomorphism, or human-likeness, exhibited through interactions with humans. The famous "uncanny valley" effect describes the feeling of eeriness and discomfort that users experience when dealing with a technology that is way too human-like [266]. In both of

my studies, I found that AI agents' human-likeness could sift through many aspects of the agent design and potentially elicit discomfort among users. During the "Design Your Agent" activity in study 2, the biggest complaint that participants had about some of the AI agents I presented was when some agents "pretend to be human" by expressing the agents' own preferences or feelings. For example, one of the AI agent dialogues I showed the participants involved an agent that could make small talks (Agent number 3 in Figure 4.6). During the dialogue the agent expressed its own preferences on the different holidays and said "I prefer to celebrate Christmas because it is a beautiful day." When discussing their likes and dislikes about Agent number 3 (as seen in Figure 4.6), P23 said, *"It made me uncomfortable when the AI was like 'I have this opinion'. I don't think you do. If it's an AI then they can't feel or have an opinion."*

Other features that triggered online students' discomfort included the use of emoji, human avatars, and human names on the AI agents themselves. Many participants explicitly expressed their preferences of agent avatar over agent number 3 and number 4 in Figure 4.6 where the agent avatars were just the three letters "BOT" (P6, P12, P14, P15, P21, P20, P23). All participants preferred AI agents to explicitly say that they were agents and not trying to be human by using human photos as avatars or human names as their names.

Social Etiquette. While the participants did not want the AI agents to exhibit human-like characteristics, I found that the agents were expected to follow human social etiquette when interacting with the students. Participants pointed out that AI agents should always ask people's feelings first before making assumptions about the users. For example, in the "Design Your Agent" activity in study 2, P19 commented on agent number 1 (see Figure 4.6) and said, *"I liked that the agent checked in with the user. But I didn't like that the agent kind of had an assumption about the user feeling ignored."* P6 also said agent number 1 made assumptions that the user didn't like being in that situation: *"It sounds a little condescending, what if you didn't want to be involved in that social situation. And you are happy about it."* Other participants also said that *"It felt like a one-sided conversation without much input*

from the user. " (P20, 22, 23) P23 suggested that this could be resolved by having the agent asking more questions to gain enough context from the user.

Another instance of AI agents violating social etiquette was when the agent interjected in the middle of a group conversation and called people out. During the discussion on agent number 4 (see Figure 4.6), many participants appreciated that the agent was trying to encourage group participation by asking opinions from students who hadn't expressed their thoughts in the group chat. However, some participants also pointed out that the AI agent calling people out by name in the group chat might make them feel uncomfortable (P16-P18). When asked to elaborate on that, P17 compared it to her experience interacting with other students, *"I remember there was one time a student said something a little rude in the group chat. I just privately messaged him to say that was not cool. But some people called him out and it even made me feel uncomfortable."* P18 agreed and suggested that, *"A better way to handle this would be direct message the students instead of calling them out in the group chat."*

Some AI agents were also perceived as "self-centered" or "patronizing" because the agent either ignored the user's message or sounded bossy. For instance, for agent number 3 dialogue (see Figure 4.6), many participants believed that the agent ignored the user's message. P18 said, *"This one seemed to be ignoring what the user said... they are just talking about themselves"* This was further echoed by P20, P22, and P23. P6 also said, *"I wouldn't reply back at all. The agent's response doesn't progress the conversation forward."* When talking about agent number 2 (see Figure 4.6), some participants found the agent's suggestion to be patronizing. P21 said, *"I feel like it's almost patronizing at the end saying 'You should...' It's like someone is telling me to do something. It's off-putting for me. Especially it's a bot, like I know you don't care."*

Intelligence. Students' desire for highly intelligent AI agents mostly centered around the agents' conversational intelligence in language comprehension. One example of this was the AI agent's ability to infer implicit information from the conversations. For example,

many participants expressed their preferences for agent number 2 (see Figure 4.6) in the "Design Your Agent" activity. P14 said, *"I did like that the counselor was able to draw some information that wasn't explicitly given from the user."* P12 also said she liked that the agent was able to infer the user was under a lot of stress because of the deadline. When talking to AI agents, participants want the agents to exhibit human-level conversational intelligence. P13 said, *"I liked it (that the agent was able to infer context) because when you're talking to a real person, that's what they do. If you told your friend that I haven't been able to sleep, and your friends know the context in your life and they say, oh that's because you've been working too hard preparing for the exam ... That's how you talk to a real person. Whereas with a lot of bot they just keep asking you your order number a hundred times and you can't really get anything out of it. "*

Participants also preferred that the agent could comprehend free text instead of pre-set answer choices. In the "Design Your Agent" activity in agent number 1 dialogue, the user only answered with "Yes." and that there were pre-set answer choices for the users to choose from. While some participants found this to be convenient and straightforward, other participants also believed that this made the interaction seem "unnatural" and that they would prefer to communicate freely with the CA like communicating with other humans.

Study 2 Findings: Social and Ethical Concerns of SAMI

Privacy. Participants displayed conflicted feelings surrounding the use of student data. On one hand, students were concerned about the continuous and large-scale data collection that SAMI demonstrated— all the challenge teams raised concerns on data privacy, that students might not feel comfortable having SAMI reading all their data (T6, T8, T10) and that students might stop asking questions on the discussion forum due to SAMI's continuous monitoring (T8, T10). On the other hand, a highly personalized agent-mediated social interaction experience was also desired by the students— the desired characteristics and functionalities of SAMI that the solution team brainstormed all highlighted SAMI's po-

tential capabilities to know more about the students through large-scale data collection on students' degree progress (T7), students' course schedule (T7), students' preferences and availability for social meet ups (T9), students' postings on the discussion forum (T9), and students' location data for in-person meet ups (T9).

During the challenge cards competition rounds, online learners further discussed their concerns around data privacy and offered some potential solutions to address these concerns. For example, in session 1, both T6 and T8 posted the challenge that students may not feel comfortable with SAMI reading all of students' data within the online program. T5 proposed a solution accordingly to mitigate students' concerns by offering explicit consent, opt-in/out process for students, and that students should be able to control what data SAMI could have access to the entire time. T7's solution to this issue was similar, by proposing that the data SAMI collected stay locally and that proper measures such as two-factor authentication to be implemented to protect sutdents' privacy. However, after some discussion, students also agreed that these measures were not the perfect solutions to the data privacy challenge. As P14 accurately put, "Informed consents are good at establishing legal distance but not good at establishing user trust." In session 3, T10 posted the challenge that, given SAMI had access to students' private conversations on the discussion forum, it might stop students from reaching out to the instructors. T9 proposed that students could easily opt in and out of what kind of data to share with SAMI. However, when we followed up and asked if the participants thought opting-in and out would be sufficient to protect students' privacy, participants also acknowledged that it might not. P22 said, *"Even something is labeled as 'anonymous', nothing is truly anonymous these days. But proper anonymization for all of students' data and that SAMI could remove students' data according to students' request would help mitigate privacy concerns."*

Emotional Burden. While SAMI's goal was to reduce students' emotional burden and isolation by connecting students with others, some students pointed out that the use of SAMI might counter its goal, and instead add on to students' emotional burden. For exam-

ple, during their brainstorming sessions of the potential challenges SAMI could have, T6 pointed out that students might feel embarrassed if they had to use SAMI to make friends. This point was echoed by other participants, saying that students might feel like they were incapable of basic social interaction and that they even needed an AI agent to help them do that. T10 also pointed out that SAMI could also hurt students' feelings unintentionally when trying to initiate interaction (P10). Specifically, in the SAMI mockup storyboard that I created, SAMI reached out to the student Lisa and said "I am sensing that you are feeling lonely..." Participants pointed out that telling students they were lonely might make students feel uncomfortable.

In the challenge cards competition and the debriefing sessions later, participants further discussed the issue of emotional burden, specifically when system transparency could add on to students' emotional burden. I discussed with the student that in the mockup when SAMI initiate the interaction with the student, I felt like it was necessary for SAMI to explain why the interaction was initiated and hence SAMI started the conversation by saying "I am sensing that you are feeling lonely... " Participants said that offering transparency into why SAMI initiated the interaction was desired, however, it would be better if SAMI could stick to factual language and avoid using emotional words like "lonely." P14 said, *"When putting it in particular terms it might risk people interpreting it in the wrong way and add to students' emotional burden."* P17 also agreed that wording would be important to avoid adding emotional burden to students: *"If SAMI said 'you haven't checked in with your classmates for a while, would you want to check in?', that would be much better than 'are you lonely?"'* As P23 summarized, the way SAMI communicated messages should be based on facts instead of trying to convey the idea that "the machine understands you."

Misinterpretation by SAMI. Another set of ethical and social challenge that online learners brought up was the possibility of SAMI misinterpreting students' social needs or preferences. Part of this concern stemmed from participants' uncertainty about how accurate were SAMI's inferences made from students' online digital footprints. For example, participants

were concerned about SAMI misunderstanding a student's social needs (T6), misinterpreting students' "emotional level" (T6, T10), or misconceiving students' level of desire for social connections (T8). Participants further elaborated on their concerns about SAMI misunderstanding their emotional level. T6 pointed out that emotions do not change linearly and could fluctuate frequently. Inferring students' emotion from their online posts on the discussion forum might miss the time that students actually need help. T10 was also doubtful that SAMI could infer students' emotion level just based on students' discussion posts which were often centered around class assignments.

To resolve this challenge, participants offered several strategies that could improve SAMI's accuracy in making inferences about the students. Both T5 and T7 proposed that SAMI should always check with the student to confirm the accuracy of SAMI's inferences made and that students should have the ability to correct SAMI's inferences if they were inaccurate. T7 also pointed out that students' social needs and preferences tend to change over time. In order to improve the accuracy of SAMI, SAMI could continuously keep track of students and adjust the inferences accordingly.

However, even if SAMI could make highly accurate inferences from students' online footprints, other challenges remain such as the malicious usage of SAMI. Both T6 and T10 believed that people's online profile could be completely different than their actual persona. Online persona could also be easily manipulated to achieve individual goals. One example that T6 gave was that if SAMI could provide more academic help when students were frustrated with the course material, some students might take advantage of this and intentionally present themselves as frustrated in order to gain more help on their assignments. Another rather extreme example given by T6 was that this could also be leveraged by cyberbullies. A hypothetical example would be that a bully would pretend to be the same type of people that they tend to bully in order for SAMI to connect them to the group of potential victims that they could bully.

4.2.6 Discussion

These findings offer insights into online learners' desired functionalities and social characteristics of AI agents that can promote social connectedness among online learners. Specifically, I identified online learners' desire for in-situ agent-mediated social support through continuous monitoring as well as the need for AI agents to scaffold their social interaction process. Based on these findings, I summarized and distilled several design implications for AI agents' functionalities and social characteristics from a human-centered perspective in Table 4.5.

I also identified online learners' discomfort about the AI agents could be triggered through agents' expressions of opinions or preferences. Online learners also wanted the AI agent to follow social etiquette and be aware of the context during interactions. These findings also shed light on the perceived social and ethical challenges of using AI agents to mediate social interactions among online learners, including concerns about privacy, emotional burden, and misinterpretation. While students' concerns of privacy and misinterpretation of their online learning data have also been suggested by prior work in learning analytics [210, 212], collecting and analyzing online learners' data for social purposes presents new unique challenges on students' perception of privacy and the possibility of AI agents adding emotional burden during agent-mediated social interaction.

The present study raises the concern that when data collection and analysis by AI serves social purposes, students could be more open and inclined to share their data in agent-mediated social interaction comparing to data collection and analysis for learning purposes. Even though the large-scale data collection and monitoring of the AI agent raised online learners' privacy concerns [216] during my study, having a highly personalized agent-mediated social interaction experience was also highly desired by online students. Prior research in social matching [31] also suggests that users are often more willing to sacrifice their data privacy to gain more accurate and personalized social matches due to humans' inherent social nature [31]. However, what online learners did not show concerns about was

Table 4.5: This table summarizes the design implications based on our findings on online learners' desired functionalities and social characteristics of AI agents that can help online learners feel socially connected. We also list examples of how to implement each design implications.

Categories	**Design Implications**	**Examples**
Functionalities	Help online students identify other like-minded students to connect.	Identify and connect students who are interested in reading, or students who are struggling with the same assignment.
	Offer continuous support to online students' changing social needs.	Continuously update online students' social needs and preferences by checking with the students or monitoring their online activities.
	Initiate interaction with online students at the right time.	Incorporate advanced AI techniques to make inferences about students' real-time status (e.g., emotional state, loneliness)
	Scaffold online students' social interaction process throughout.	Introduce online students together with ice-breaker questions or schedule meetups and social events based on student schedule.
Social Characteristics	Having personality is not necessary.	Personality could be neutral without overly expressing humors or emotions.
	Use non-humanlike avatars and names.	A simple "BOT" avatar could convey clearly that the agent is not a human.
	Don't pretend to be humans.	AI agents should avoid saying things like "I prefer to celebrate Christmas." which implies that the agent has opinions or preferences.
	Follow social etiquette during conversations.	AI agents should not ignore users or use seemingly condescending language such as "you should do X."
	High conversational intelligence in language comprehension.	Understand the context of current conversations through cues by free text.

the fact that the AI agent also collects and analyzes data from private conversations between the AI agent and individual students. Existing literature has pointed out that AI agents could elicit more private and sensitive personal information (e.g., credit card information [102]) from the users comparing to humans [217, 102, 218]. This is due to people's "general machine heuristics" which refer to people's rule of thumb that AI agents as machines, when compared to humans, are perceived as more trustworthy and secure [102]. Combining with the social purposes of agent-mediated social interaction, AI agents' capability of encouraging deep self-disclosure, and students' heightened trust in schools collecting and using their data [15], a concern is that online students could be put in an even more vulnerable position to have their data exploited in agent-mediated social interactions.

Modest concerns regarding the communication of learning analytics results to students have been raised in prior literature. However, the social nature of agent-mediated social interaction poses new social challenges of AI agents adding to online learners' emotional burden when they are socially isolated. In my study, online learners pointed out that the AI agent's emotional language to convey system transparency in initiating the conversation with students, "I am sensing that you are feeling lonely..." could add on to students' emotional burden. While AI agents communicating in emotional and judgemental language could cause harm to users has been hypothesized in prior literature [217], my findings provide empirical evidence to validate this concern. Therefore, designers should be mindful of the languages that AI agents use when communicating with students who could be feeling socially isolated to avoid adding further emotional burden on them. While communicating transparency about AI agents' working mechanism is highly desired [233, 235], AI agents should stick to factual, unbiased, and non-judgmental language [217, 245] when communicating with students. For example, if students' emotional states were monitored and used to initiate interactions, AI agents should avoid using emotional labels such as "I am sensing that you are lonely..." but could explain the initiation by communicating analytics-based language like, "I noticed you have lower forum activity than normal..."

4.2.7 Limitations and Future Research

While this work offers important design and ethical implications on using AI agents to promote online learners' social connectedness, this work has several limitations. First, this study assumed that AI agent, or even technology in general, could help improve online learners' social connectedness. This assumption could have influenced participants' ideas of what is possible to help reduce their feelings of social isolation. Second, the perceived social ethical challenges of agent-mediated social interaction were elicited based on one AI agent mockup that I created. I acknowledge that there could be many ways in which AI agents can facilitate and scaffold online learners' social interaction and we encourage future research to further examine the social and ethical concerns in other types of agent-mediated social interaction. Third, all participants were recruited from the OMSCS program at Georgia Tech. Therefore participants in our study could have more, or less, concerns compared to online students in other disciplines that are less technology-centered. Future work should replicate our study with online students in less technology-centered disciplines (e.g., liberal arts) to gather a complete range of concerns that online learners might have about agent-mediated social interaction. Finally, our findings might not be applicable to other forms of online learning environment such as Massive Online Open Classes (MOOC) or online learning programs at undergraduate or K-12 level.

4.3 Reflections & Takeaways

In this chapter, I described two studies conducted to understand the human-centered design of AI-mediated social interaction in the context of online learning, where AI-mediated social interaction is urgently needed. Through these studies, I identified online learners' current practices and challenges in building remote social connections in online learning such as the lack of visibility of social signals, diminished awareness of potential social companions, decreased accountability in social behaviors, and the lack of spontaneity and

randomness in online learning environment. I discussed design opportunities for AI systems to mitigate these challenges through promoting social translucence and bridging the social-technical gap, highlighting the promising potential of using human-like AI agents to mediate remote social interactions. Based on these findings, I then conducted co-design workshops with online learners to derive a set of design guidelines that detailed the desired functionalities, social characteristics, and ethical concerns of using AI agents to mediate remote social interactions.

Through qualitative and design approaches, these two studies took online learning as an exemplar of large-scale learning context to understand the design of AI-mediated social interaction through online learners' perspectives. However, both of these studies were conducted with OMSCS students at Georgia Tech, which presents certain limitations on the design implications. While efforts were spent on balancing the participants' gender and age diversity in each study, the OMSCS students are computer science students who are more tech-savvy and have more positive view on AI technologies than average online students. Findings from these studies are also more applicable to online for-degree programs with structured courses like the OMSCS program compared to other online learning environment such as MOOCs. Future work should replicate these studies with students from non-CS majors and online learning environment with less structures to provide design implications that could be transferable to other online learning contexts.

The findings from these two studies highlight the tension between AI design *from* the students and AI design *for* the students. The first study underlined online learners' preferences for human-like AI agents to mitigate the social-technical gap in remote social interactions by providing naturalness and spontaneity; yet the second study surfaced online learners' concerns for such human-like AI agents that could present privacy and even emotional harms due to agents' human-like characteristics and functionalities. This tension underlines the need for future human-AI interaction to maximize the benefit of AI agent's human-like features while mitigating potential harms stemmed from perceived AI

anthropomorphism. This highlights the design direction for human-AI communication to carefully manage and account for people's perceptions of AI. Inspired by human's ToM capability to manage their impressions of each other through MToM during human-human communications, the rest of this thesis presents three studies, each corresponding to one stage of the MToM framework, to understand the ToM construction, recognition, and revision process of both the human's and the AI's iterative interpretations of each other. Next chapter begins this series of empirical exploration by presenting a study that examines the first MToM stage— ToM construction— through investigating online learners' longitudinal perceptions of a virtual teaching assistant.

CHAPTER 5

TOM CONSTRUCTION: AI'S CONSTRUCTION OF HUMAN'S INTERPRETATION OF THE AI

This chapter explores the first stage of the MToM framework for human-AI communication in AI-mediated social interaction: *ToM Construction: AI's construction of human's interpretation of the AI*.

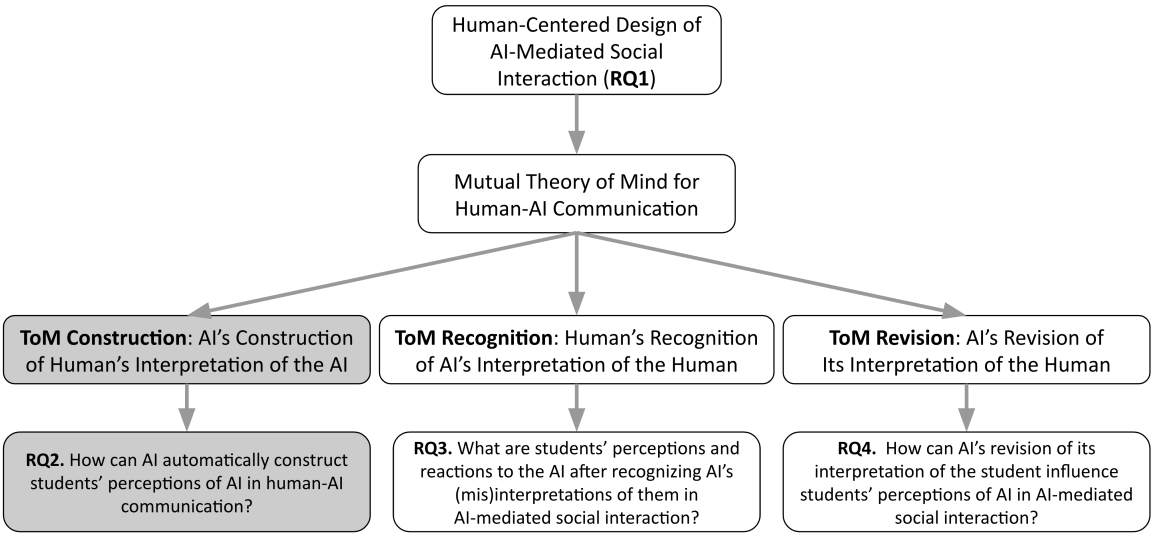


Figure 5.1: Chapter 5 explores ToM construction: AI's construction of human's interpretation of the AI.

The ToM construction stage describes the ToM process of AI constructing human's interpretation of the AI based on human's communication feedback. ToM construction is the fundamental stage of achieving MToM in human-AI communication. An AI system capable of automatically inferring and constructing human's perceptions of the AI would allow the AI system to generate communication feedback that can intentionally shape people's perception of the AI. This would also help alleviate the current one-sided communication burden on users, who had to constantly adjust their mental model of the AI system through an arbitrary trial-and-error process to elicit desired responses [142, 22]. In this chapter, I

explore the following research question:

RQ2. How can AI automatically construct students' perceptions of AI in human-AI communication?

Constructing accurate human interpretation of the AI requires the AI system to extract valid communication cues from human's communication feedback. Human's communication feedback can take many forms such as language, gestures, facial expressions, etc. However, most of these communication cues are lost due to the asynchronous nature of large-scale learning environment, especially online learning. Textual communication is the most prevalent form of communication in online learning environment. Examining valid linguistic cues that can reflect students' perceptions of the AI from students' communication feedback is thus at the core of the ToM construction stage. Constructing student's perception of the AI through linguistic cues would enable AI agent's ToM capability as shown in Figure 1.2 in chapter 1, where the AI agent can automatically detect the student's inaccurate perception of the AI agent and respond accordingly to correct their misinterpretation of the AI.

To understand this problem, this chapter presents a longitudinal survey study that examined students' long-term perceptions and textual communications with a question-answering AI agent deployed as a virtual teaching assistant in an online class. Virtual teaching assistant, similar to AI social facilitator SAMI, plays a critical social role in the online student community to provide informational support to individual students. The question-answering functionality of such virtual teaching assistants provide abundant amount of textual information to conduct experiments to extract valid linguistic signals that can predict student perceptions. Through longitudinal surveys and linguistic analysis, I found that students' perceptions of the AI agent fluctuated significantly over time, even when the AI agent did not exhibit any learning capability. I also found that linguistic characteristics of students' utterances to the AI agent can reflect their perceptions, specifically likeability, intelligence, and anthropomorphism, of the AI agent. This study established the feasibility

of equipping AI systems with the ToM-like capability to automatically construct students' changing perceptions of the AI through language analysis.

5.1 Introduction

Conversational Agents (CAs)[1] are becoming increasingly integrated into various aspects of our lives, providing services across healthcare, entertainment, retail, and education. While CAs are relatively successful in task-oriented interactions [267, 268], the initial promise of building CAs that can carry out natural and coherent conversations with users has largely remained unfulfilled due to both design and technical challenges [269, 270, 271]. This "gulf" between user expectation and experience with CAs [272] has led to constant user frustration, frequent conversation breakdowns, and eventual abandonment of CAs [272, 273, 271].

Conducting smooth conversations with users becomes even more crucial when CAs are deployed in online communities, especially those catering to vulnerable populations such as online health support groups [274] and student communities [275]. These community-facing CAs often serve as a critical part of the community to ensure smooth interactions among community members and provide long-term informational and emotional support. However, these community-facing CAs face two unique challenges: the need to carry out smooth dyadic interactions with individual community members, and the need to respond accordingly based on the community's shifting perceptions [276, 277]. In fact, the community-facing nature of the CA adds new complexity— each dyadic interaction with individual members is visible to other community members, which can not only change the community's perception of the CA, but can also impact other community members, i.e., unsatisfactory interaction with one individual might also frustrate others [278].

Inspired by the MToM framework, I posit that equipping CAs with an analog of ToM that can automatically identify user perceptions about the CAs would improve human-AI

[1]Unless indicated otherwise, I use CAs to refer specifically to disembodied, text-based conversational agents.

communication process as well as user experience. While research has explored ways along the realm of identifying user perceptions of CAs to facilitate dyadic human-AI interactions, these studies, most of which are qualitative in nature, are not only difficult to scale, but also lack directly feasible algorithmic outcomes that could be integrated into CA architecture to automatically recognize user perception about the CA. For community-facing CAs that are known to have fluid social roles in online communities [270], we presently lack a clear understanding of how community perception of CAs evolve over time, and whether the very dyadic interactions between humans and CAs in community settings reveal any signal related to user perceptions.

I thus note a gap in theory and practice in automatically and scalably understanding human perceptions of a community-facing CAs at both individual and collective level. Drawing on the dynamics of human-human interactions, this paper explores a first step towards designing for MToM in long-term human-CA interactions by examining the feasibility of building community-facing CAs' ToM. Specifically, I target two research questions in this study:

RQ 2.1: How does a community's perception of a community-facing CA change over time?

RQ 2.2: How do linguistic markers of human-AI interaction reflect perception about the community-facing CA?

I examine these research questions within the context of online learning, where community-facing CAs are commonly seen to provide informational and social support to student communities [279, 275, 35]. I deployed a community-facing question-answering (QA)CA named **Jill Watson** [40, 280, 281] (JW for short) in an online class discussion forum to answer students' questions for 10 weeks over the course of a semester. I collected students' bi-weekly self-reported perceptions and conversations with JW for further analysis. I discuss changes in the student community's long-term perception of JW and examine the

relationship between self-reported student perceptions of JW and linguistic attributes of student-JW conversations such as verbosity, adaptability, diversity, and readability. Regression analyses between linguistic attributes and student perceptions of JW reveal insightful findings such as readability, sentiment, diversity and adaptability positively vary with desirable perceptions, whereas verbosity varies negatively.

5.2 Study Design

5.2.1 Study Overview

Current study seeks to understand longitudinal changes of community perception of a community-facing CA and the feasibility of leveraging linguistic markers to infer user perceptions of a community-facing CA. To explore these questions, I deployed a community-facing (QA)CA named "Jill Watson (JW)" in an online class discussion forum to answer students' class logistic questions throughout the semester. I then collected students' bi-weekly perceptions of JW and extracted linguistic characteristics from student-JW conversations over the course of the semester (see Figure 5.2 for detailed study design). I selected our survey measures and linguistic features with the goal of ultimately building CA's ToM— survey measures were designed to gauge students' perception of JW from three dimensions: anthropomorphism, intelligence, and likeability; linguistic characteristics were suggested by prior literature to have the potential of reflecting people's perception of the CA. I discuss them in detail in the following sections.

5.2.2 Design and Implementation of JW

JW is an ML-based question-answering CA designed to answer students' questions about class logistics. It uses three machine learning models with each model being trained with the same data. When a user asks a question, the question is passed to all three models. The final output of the models is used to select a pre-programmed response (greetings + relevant information in the syllabus). The models were trained using training questions generated

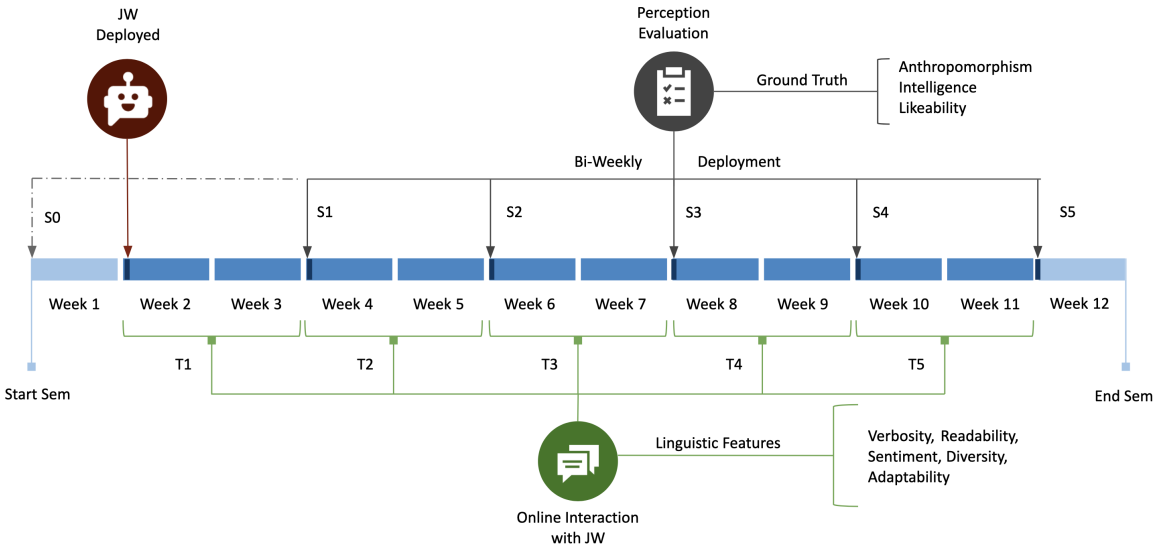


Figure 5.2: Study design and timeline. S0-S5 represents the survey data. T1-T5 represents our division of class discussion forum data based on the survey distribution timeline. In the regression analysis, I used survey data as ground truth to tag student interaction with JW in each time frame. For instance, I used S1 to tag forum data from T1, S2 to tag T2, and so on.

from a knowledge base. The knowledge base was created using a syllabus ontology and the course syllabus. JW thus *cannot* learn from outside information (student responses or feedback) over time. Implementation details of similar previous versions of JW can be found in Goel and Polepeddi (2016) [40].

I deployed JW on the class discussion forum at the beginning of the second week (Figure 5.2). JW was only active on dedicated "JW threads" where JW read and provided responses to each post to only questions posted in those threads. Students were encouraged to post their class-related questions on this thread if they wanted an answer from JW. To keep students engaged throughout the semester, I posted a new "JW thread" every week on the discussion forum and encouraged students to keep asking questions to JW. Table 5.1 shows a list of example question-answer pairs between the students and JW on the class discussion forum.

Throughout my study, *I intentionally did not specify JW's working mechanism or capabilities to the students* so that the information would not bias the students' perception of

Table 5.1: Examples of question-answer pairs during students' interactions with JW throughout the semester on the class discussion forum thread.

Week	Student Question	JW Answer
Week 4	What is the next assignment?	Hi, happy to answer your question! There are ten written assignments in the course: five P assignments and five M assignments. The P assignments are Principles assignments and cover the material covered in Unit 2. Each Principles assignment asks you to [...]. The M assignments are Methods assignments and cover the material in Unit 3. Each Methods assignment asks for [...].
Week 4	How were you programmed?	Hi, happy to answer your question! Sorry I don't understand your question. Could you rephrase your question by composing a new followup discussion?
Week 6	Hi Jill Watson, is the exam going to cover the concepts we applied in the assignments?	Hi, happy to answer your question! Here are some questions that I think you might be asking. Please compose a new followup discussion with one of the questions here: "Can I work ahead?", "What is the JDF specification?" "Can we look at online resources?"
Week 8	What's the most efficient way to earn participation points?	Hi, happy to answer your question! HCI is a deeply collaborative field, and there is no better way to learn than to interact with your peers, to participate in one another's usability studies, and to see the variety of approaches taken by your classmates to the class's assignments. Thus, participation credit may be earned in one of three ways: by completing peer reviews, by participating in one another's surveys and interviews, and by participating in discussions on Piazza. The following are the ways you may earn participation points [...]
Week 10	When is the mid-course survey due?	Hi, happy to answer your question! Please complete this survey by the end of week [link to mid-course survey]. You receive 1 participation points for completing one of the four course surveys. For more details, see the participation policy [...]
Week 12	Hi Jill Watson, could you predict the percentage of students who would get an "A" at the end of class?	Hi, happy to answer your question! Please complete this survey by the end of week [link to mid-course survey]. You receive 1 participation points for completing one of the four course surveys. For more details, see the participation policy [...]

JW. The students were only told that JW was a virtual agent who could answer their questions about the class. JW's working mechanism and implementation were only revealed after all the survey data was collected.

5.3 Examining Changes in Student Perceptions about the AI Agent

5.3.1 Data Analysis

To explore changes in students' perceptions of JW throughout the semester, I deployed six bi-weekly surveys (See Appendix B for the adapted survey instrument) for students to self-report their perceptions of JW. Inspired by the MToM theoretical framework, I intentionally selected perception metrics that could capture students' holistic social perceptions of JW and potentially reflect long-term changes in perceptions of JW, instead of the commonly measured post-hoc perceptions of CA functionalities (e.g., accuracy or correctness of response).

In particular, I adapted a validated survey instrument measuring user perception of robots in human-robot interactions [282], also previously applied in human-CA interaction settings [283]. In my specific setting of student-JW interactions, my surveys inquired students to self-report their perceptions about JW along three dimensions: 1) anthropomorphism, 2) intelligence, and 3) likeability. In addition, I also asked the students to report how/if they interacted with JW in the past two weeks (e.g., read other students' interactions with JW, posted questions to JW).

Data. I started with an initial total dataset of 1513 responses from $\mathtt{S}_0$ to $\mathtt{S}_5$. I consolidated all our responses to build our final dataset that included all valid, complete responses from students who indicated that they interacted with JW by either reading through other students' interactions or posting questions to JW. I ended up with a total of 1132 responses from $\mathtt{S}_0$ to $\mathtt{S}_5$ ($N_{S0} = 260, N_{S1} = 201, N_{S2} = 171, N_{S3} = 171, N_{S4} = 164, N_{S5} = 165$).

My analyses did not include $\mathtt{S}_0$ survey results that indicate students' expectation of

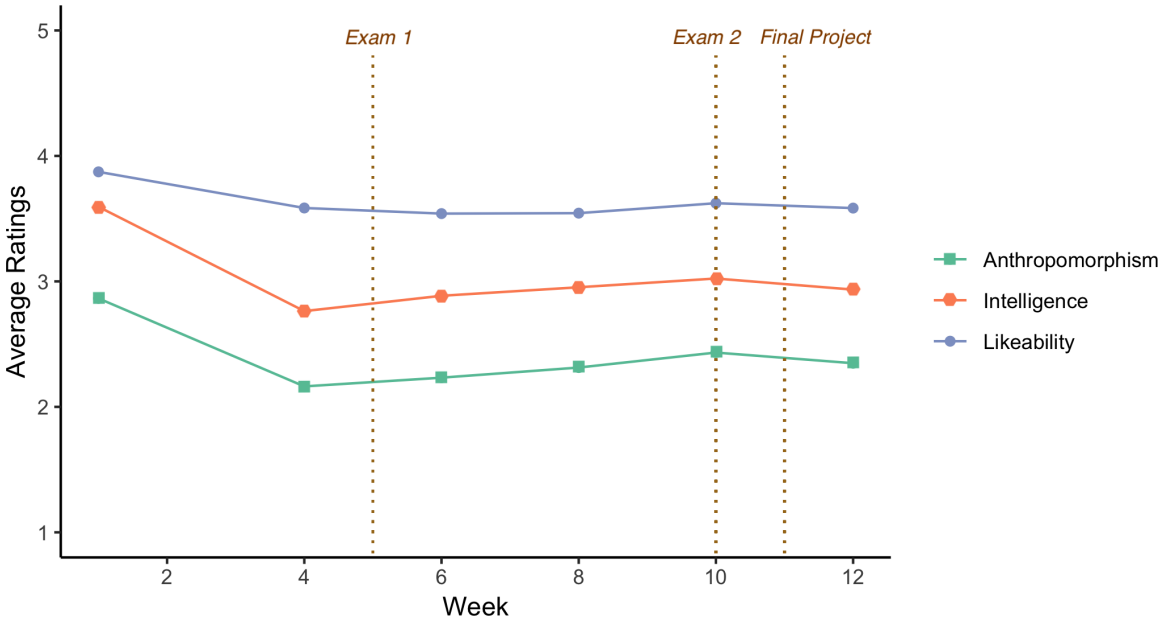


Figure 5.3: Student perceptions of JW over time. To provide more context, the plot marks the due dates of Exam 1, Exam 2, and Final Project. Note that students in this class also have weekly written assignments.

JW prior to actual interactions as I am more interested in examining long-term changes in student perception *after* at least some initial interaction. It is also well-established in the literature that people often have unrealistically high expectations for CAs [272, 273]. My findings replicate this similar pattern from prior literature, that students' perception *decreased* compared to their initial perception (or expectation) as per S_0, as is revealed in Figure 5.3, which plots the aggregated community perception about JW over the course of the semester.

5.3.2 Findings

Next, to understand if student perception of JW changed significantly after initial interactions, I performed Kruskal-Wallis test [284] on students' self-reported perception of JW from S_1 to S_5. Kruskal-Wallis test is a non-parametric, omnibus test, which I used because the data is not normally distributed based on the results of a Shapiro-Wilk normality test ($p < 0.001$). I then conducted further post-hoc pairwise comparison to examine differences

Table 5.2: Summary of comparison in students' bi-weekly perceptions of JW. I report Kruskal-Wallis test results for each perception metrics from S_1 to S_5, the posthoc pairwise comparison z statistic (Dunn Test), and effect size (Cohen's d). p-values are reported after Bonferroni correction (* p<0.05, ** p<0.01).

	Anthropomorphism		Intelligence		Likeability	
Measure	***z***	***d***	***z***	***d***	***z***	***d***
S_1 and S_2	-0.60	0.08	-1.63	0.16	0.67	0.06
S_1 and S_3	-1.47	0.17	-2.32	0.25	0.69	0.06
S_1 and S_4	-2.82*	0.31	-3.26**	0.33	-0.59	0.05
S_1 and S_5	-1.88	0.21	-2.13	0.22	0.04	0.00
S_2 and S_3	-0.83	0.10	-0.66	0.09	0.02	0.00
S_2 and S_4	-2.14	0.22	-1.59	0.18	-1.20	0.11
S_2 and S_5	-1.23	0.13	-0.49	0.06	-0.60	0.06
S_3 and S_4	-1.32	0.13	-0.93	0.09	-1.22	0.11
S_3 and S_5	-0.41	0.04	0.16	0.02	-0.62	0.06
S_4 and S_5	0.90	0.10	1.09	0.11	0.60	0.05
Kruskal-Wallis	$\chi^2(4) = 9.55 * *$		$\chi^2(4) = 11.81*$		$\chi^2(4) = 2.09$	

between each bi-weekly perception report. Dunn Test result shows significant differences in perceived anthropomorphism between S_1 and S_4: $z = -2.82, p = 0.02$, and significant differences in perceived intelligence between S_1 and S_4: $z = -3.26, p = 0.01$. I reported the detail test results and effect sizes in Table 5.2.

Anthropomorphism. Anthropomorphism is the attribution of human characteristics to non-human objects such as computers and CAs. Anthropomorphism is a widely studied yet highly debatable design characteristic of CA—on one hand, intentionally building CAs with more humanlike attributes can improve user trust[285, 286], make the CA more approachable and ease user interactions [287, 288]; on the other hand, the famous "Uncanny Valley" effect [289] indicates that highly anthropomorphized CA could evoke people's negative feelings towards the CA [266] as well as setting unrealistic user expectations on CA's capabilities [272]. Changes in perceived anthropomorphism over time is thus an important quality to investigate as it could significantly affect people's expectation of the CA and thus influence trust-building and long-term human-agent relationship [290]. Kruskal-Wallis test found students' self-reported perceived anthropomorphism after initial interaction with JW *changed significantly* over time from S_1 through S_5: $\chi^2(4) = 9.55, p < 0.05$. Post-hoc pair-

wise comparison found $\texttt{S}_1$ and $\texttt{S}_4$ differ significantly: $z = -2.82, p < 0.05$. This indicates that CAs' perceived humanlikeness by the community can vary over time, even when the agent has zero learning ability and adaptability.

Intelligence. Intelligence refers to the perceived level of intelligence of the CA by the community, in other words, how much users perceive the CA as an intelligent being. Even though building artificially "intelligent" machines has been a unfulfilled promise due to various technical and feasibility challenges [270, 282], users tend to expect their CAs to be "smart" [273], thus creating a gap between user expectation and CA's true intelligence. CA's knowledge is also one of the key components identified in people's mental model of CA [96]. Therefore, perceived intelligence plays an important role in how people perceive, evaluate, and interact with the agent. However, it is unclear whether people's perception about the CA's intelligence change over time. In this study, Kruskal-Wallis test found that the perceived intelligence of JW changed significantly from $\texttt{S}_1$ to $\texttt{S}_5$: $\chi^2(4) = 11.811, p < 0.05$, specifically, post-hoc pairwise comparison shows that perceived intelligence reported in $\texttt{S}_1$ and $\texttt{S}_4$ differ significantly: $z = -3.26, p < 0.01$. This highlights a CA's perceived intelligence is an important attribute to consider when building long-term human-AI relationships.

Likeability. Likeability refers to how likeable the interlocutor is perceived by others. In human interactions, likeability has been suggested to induce positive affect, increase persuasiveness, and foster favorable perceptions [291, 292]. Since people often treat computers as social actors [93, 282], perceived likeability is a potential factor that could influence long-term relationship-building. Kruskal-Wallis test could not find statistically significant changes in students' self-reported likeability of JW over time: $\chi^2(4) = 2.0947, p = 0.72$. This result could be attributed to the fact that positive first impression in human interactions typically plays a crucial role in long-term likeability [293]. Another reason could be that students initial perception of JW remains the same over time since JW was intentionally

designed to be a basic CA without learning ability.

Correlation Between Perception Measures. I also conducted Spearman correlation test, a non-parametric correlation test, to examine the relationship between these three perception measures. Spearman correlation results show that perceived anthropomorphism and intelligence have a strong positive relationship ($r_s = (0.74)$, $p < 0.001$), intelligence and likeability have a moderately strong positive realtionship ($r_s = (0.62)$, $p < 0.001$), and anthropomorphism and likeability have a low positive correlation ($r_s = (0.51)$, $p < 0.001$). This result suggests that even though the three measures of perception are considered to be somewhat independent [282], that may not be the case in my data. That is, according to my data, students' perception of desirability along the three measures are in similar direction, a general increasing trend in one would likely convey in a general increasing trend in the other two.

Summary and Interpretation. Through analyzing students' bi-weekly self-report of their perception of JW, I conclude that JW's perceived anthropomorphism and intelligence significantly changed over time, but perceived likeability did not significantly vary in the long run. These findings help me understand how community perceptions of a community-facing CAs change. This bears implications on designing community-facing CAs to be able to adapt to community's changing perceptions of the CA in the long run. I also found the three measures of self-reported perceptions to be inter-correlated, shedding light that these measures may not be very disentangled (or independent) in users' mental models.

5.4 Language Reflects Student Perceptions about the AI Agent

In this section, I examine the relationship between how the students perceived and linguistically interacted with JW. To do this, I collected the conversation logs between students and JW from all the weekly question and answering threads on the public discussion forum and then extracted linguistic features for further data analysis. With the goal of exploring

the feasibility of building a ToM for CAs, the linguistic measures were chosen due to their known potential in reflecting users' holistic perceptions of CAs, which I refer to relevant research and describe in more details in the following sections. I also discuss the findings and implications for designing human-CA interactions.

5.4.1 Data Analysis

First, I link students' linguistic interactions with JW in a block of time with their immediate next self-reported perception about JW as ground-truth. For example, if a student made multiple posts to JW from week 4 to week 6 ($\mathtt{T}_2$) and reported their perception of JW in week 6 ($\mathtt{S}_2$), then for this student, I derive language features of $\mathtt{T}_2$ to understand their self-reported perception in $\mathtt{S}_2$. Such an approach enables me to examine if the linguistic interaction between a student and JW in a block of time can predict how they would perceive the agent immediately at the end of that time block. This leads me to a total of 551 pairs of linguistic interactions and self-reported perceptions with $\mathtt{N}(\mathtt{T}_1) = 157$, $\mathtt{N}(\mathtt{T}_2) = 86$, $\mathtt{N}(\mathtt{T}_3) = 126$, $\mathtt{N}(\mathtt{T}_4) = 96$, $\mathtt{N}(\mathtt{T}_5) = 86$.

Next, I build linear regression models. Linear regression is known to help interpret conditionally monotone relationships with the dependent variable [294]. In particular, I build three linear regression models where each model uses one of the three perception measures as the dependent variable. I draw on prior research to derive a variety of linguistic attributes (features) from the language interactions which include verbosity, readability, sentiment, diversity, and adaptability [295, 296]. I use these linguistic features as independent variables in the models. As both perception and linguistic interactions could be a function of time, I include an ordinal variable of the week of the datapoint as a covariate in the models. Further, I control the models with an individual's baseline language use, particularly the baseline average number of words computed over all the posts made by the same individual. Equation 5.1 describes the linear regression models, where $\mathcal{P}$ refers to the measures of anthropomorphism, intelligence, and likeability.

Table 5.3: Coefficients of linear regression between students' perception (as dependent variable) and language based measures of interaction with JW (as independent variables). **Purple** bars represent the magnitude of positive coefficients, and **Golden** bars represent the magnitude of negative coefficients. . p<0.1, * p<0.05, ** p<0.01, *** p<0.001.

	Anthropomorphism		Intelligence		Likeability	
Measure	**Coeff.**	**p**	**Coeff.**	**p**	**Coeff.**	**p**
Baseline Avg. Num. Words	0.15	***	0.16	***	0.13	***
Week	0.06	***	0.06	***	0.03	**
Verbosity						
Num. Unique Words	-3.34	**	-3.37	**	-3.91	*
Complexity	-1.33	***	-1.82	***	-2.00	***
Readability	2.33	***	2.41	***	3.00	***
Sentiment	0.10		0.69	**	0.64	***
Linguistic Diversity	0.17	***	0.09		0.20	
Linguistic Adaptability	1.02	**	1.53	***	2.55	***
Adjusted R^2	0.85	***	0.93	***	0.95	***

$$\mathcal{P} \sim Baseline + Week + Verbosity + Readability + Sentiment + Diversity + Adaptability \tag{5.1}$$

Summary of Models. Our linear regression models reveal significance with $R^2(\texttt{Anth.}) = 0.85$, $R^2(\texttt{Intel.}) = 0.93$, $R^2(\texttt{Like.}) = 0.95$; all with $p < 0.001$. Table 5.3 summarizes the coefficients of each dependent variable. First, I note the statistical significance of the control variables, *week* and *baseline word use*. I find that people who are more expressive are more likely to have a positive perception of the agent on all three perception measures. I find verbosity to be negatively associate with each measure of perception, while adaptability, diversity, and readability positively associate with student perception of JW. Next, I describe my motivation, hypothesis, operationalization, and observation for each of the linguistic features below.

5.4.2 Findings

Verbosity. In human-human conversations, we tend to use shorter and less complex sentences when talking to a kid from sixth-grade versus when talking to an adult co-worker [297]. The verbosity of conversational language we produce thus depends on our mental model of how intelligent we perceive our interlocutor to be, which will drive the way we communicate our cognitive planning and execution of thoughts to others [296]. Translating from human-human to human-CA conversational settings, verbosity may vary on the basis of how intelligent and human-like we perceive the CA to be [298]. Hill *et al.* found that humans use less verbose and less complicated vocabulary when communicating with CAs, as compared to human-human conversations [298]. Further, the human-likeness of a CA could be judged based on the length of words used [299]. Someone who perceives a CA to be more human-like or more intelligent would likely use more verbose language. Accordingly in our setting, *I hypothesize that greater verbosity is associated with a more positive perception of JW*.

Drawing on prior work [298, 295], I use two measures to describe the verbosity of students' posts: 1) *length* and 2) *linguistic complexity*. I operationalize *length* as the number of unique words per post, and *complexity* as the average length of words per sentence [295, 296].

The regression model (Table 5.3) suggests that both verbosity attributes show negative coefficients with all the perception measures along with statistical significance. This *rejects our hypothesis*. Contrary to prior research and popular belief [298, 300, 299], my findings suggest that students who used more number of unique words per post or more complex language tended to perceive JW as less human-like, less intelligent, and less likeable. I construe that more verbose and complex language could plausibly cause the CA to fail in providing supportive or efficacious responses, leading to undesirable CA perception.

Readability. Readability refers to the level of ease readers can comprehend a given text [301]. Psycholinguistic literature values readability to be a key indicator of people's cognitive behavior, and prior work has adapted this measure to understand conversational patterns in online communities [296, 295, 302]. While this measure has not been studied in the context of human-AI interactions, from the perspective of MToM, the readability of students' questions posted to JW can convey their perception of JW's text-comprehension ability. Therefore, I examine readability to understand students' interaction with JW. However, considering an analogy from human-human to human-AI conversations, I hypothesize that, *higher readability is an indicator of a more positive perception about the CA*.

To capture the readability of students' posts to JW, I calculate the Coleman-Liau Index (CLI). CLI is a readability assessment that approximates a minimum U.S. grade level required to understand a block of text, and is calculated using the formula: $\texttt{CLI} = 0.0588\texttt{L} - 0.296\texttt{S} - 15.8$, in which L is the average number of letters per 100 words and S is the average number of sentences per 100 words [303].

The regression model shows that readability is positively associated with all three dimensions of student's perception of JW with statistical significance: anthropomorphism (2.33), perceived intelligence (2.41), and likeability (3.00). This result *supports our hypothesis*, suggesting that readability is a strong predictor of students' perception and positively varies with perception. This could be associated with an underlying intricacy that the more readable the question is, the more successful the CA response is, and the more satisfied (or positively perceiving) the users are.

Sentiment. During human-CA conversations, the emotion we convey through our language is often a manifestation of whether CA's perceived performance matches our expectations of the CA [272]. In fact, sentiment analysis has been used to detect customer satisfaction with customer service chatbots and yielded positive results [304]. Besides the perceived likeability of the CA, sentiment in the language is also positively associated with the perceived naturalness of the human-CA interactions [305, 298]. While there is a lack of evi-

dence on how sentiment in wording can be associated with perceived intelligence, intelligence is one of the key desired characteristics people expect from a CA [287]. Therefore, *I hypothesize that sentiment in students' questions posted is positively associated with a positive perception of JW*.

To measure the sentiment of each post to JW, I used the VADER sentiment analysis model [306], which is a rule-based sentiment analysis model that provides numerical scores ranging from -1 (extreme negative) to +1 (extreme positive).

The regression model (Table 5.3) shows a lack of evidence to support our hypothesis in the case of anthropomorphism, but a statistically significant support for hypothesis related to perceived intelligence (0.69) and likeability (0.64) with positive coefficients. The current study setting that JW was deployed in is considered a formal academic environment and thus themed discussion related to coursework is more common. I believe in settings where the affective language is much more prevalent (e.g., on online Reddit communities), sentiment might play a strong role in reflecting people's perception of a community-facing CA.

Linguistic Diversity. Depending on our perception of the interlocutor, the linguistic (and topical) diversity of our language could vary, i.e., the diversity of the conversation topics or the richness of language used. Linguistic diversity has been suggested to correlate with perceived intelligence during human-human interactions [297]. In human-CA interactions, when the CA behaves in a more natural and authentic way, users also tend to employ a richer set of language, conveying positive attitudes towards the CA [305]. Therefore, *I hypothesize that the greater the linguistic diversity is, the more positive students perceive JW.*

I draw on prior work [307, 295] to obtain linguistic diversity, and use word embeddings for this purpose. Word embeddings represent words as vectors in a higher dimensional latent space, where lexico-semantically similar words tend to have vectors that are closer [308, 309, 310]. In our case, for each post to JW, I first obtain its word embedding representation in 300-dimensional latent lexico-semantic vector space using pre-trained

word embeddings [309]. I then compute the average cosine distance from the centroid of all the posts by the same user in each two-week period before corresponding surveys. This operationalizes the measure of lexico-semantic diversity of each student's post to JW.

According to the regression model, I find a lack of support for my hypothesis for perceived intelligence and likeability, whereas, a statistically significant support for my hypothesis on anthropomorphism which shows a positive coefficient (0.17). This finding adds some support to previous work on human-CA interaction that suggested positive association between high lexico-semantic diversity and perceived human-likeness of the CA [305]. Contradictory to observations related to human-human interactions [297], my observations suggest that people's linguistic diversity does not necessarily indicate how intelligent one perceives an agent to be.

Adaptability. As humans, we tend to adapt to each other's language use during conversations due to our inherent desire to avoid awkwardness in social situations [30]. Prior research suggested that people often mindlessly apply social rules and etiquette to computers [311], it is thus possible that we also adapt our language when conversing with a CA. In fact, prior work suggests that we are able to adapt our speech pattern accordingly based on whether the interlocutor is a human or a CA [298], suggesting that adaptability of our speech pattern could be an indicator of our perception of interlocutor's intelligence, human-likeness, as well as likeability. Human users are more likely to build desirable perceptions about a CA if CA response is adapted and customized to human questions, as opposed to templated responses (e.g., "Thank You", "Sorry") [295]. Therefore, *I hypothesize that adaptability is positively associated to perceived anthropomorphism, likeability, and intelligence*.

Motivated by Saha and Sharma's approach [295], I measure adaptability as the lexico-semantic similarity between each question-response pairs of student-JW interactions, operationalized as the cosine similarity of word embedding representations of the questions and responses. As in the case of diversity, I use 300-dimensional word embedding space [309].

The regression model indicates that adaptability positively associates with anthropomorphism (1.02), intelligence (1.53), and likeability (2.55), all with statistical significance. This supports my hypothesis, and aligns with prior research on how people employ different speech patterns depending on if the interlocutor is a CA or a human [298]. My observations suggest that adaptability is a valid predictor of the perceptions of JW. I construe that if students receive adaptable responses, they are more likely to perceive JW as more human-like, likeable, and intelligent.

Summary and Interpretations. I examine the relationship between linguistic features of student-JW conversations and student perception of JW through regression analysis. I find that verbosity negatively associates with student perception of JW, whereas readability, sentiment, diversity, and adaptability positively associate with anthropomorphism, intelligence, and likeability. My findings suggest the potential to extract linguistic features to measure community perceptions of CA during conversation, and thus enable the CA to constantly understand and provide desirable responses that match with user perception. It is important to note that the relationship between linguistic measures and three measures of student perception of JW is of the same degree and direction.

5.5 Discussion

My findings provide empirical evidence on the long-term variations in a community's perception of a community-facing CA as well as the feasibility of inferring user perceptions of the CA through linguistic features extracted from the human-CA dialogue. Specifically, I found the student community's perception of JW's anthropomorphism and intelligence changed significantly over time, yet perceived likeability did not change significantly. The regression analyses reveal that linguistic features such as verbosity, readability, sentiment, diversity, and adaptability are valid indicators of the community's perceptions of JW. Based on these findings, I discuss the implications of leveraging language analysis to facilitate

human-AI interactions. Then, I present the challenges and opportunities for designing adaptive community-facing CAs.

5.5.1 Language Analysis to Design Human-AI Interactions

This work demonstrates that leveraging linguistic features extracted from human-CA conversations has the potential to improve human-CA interactions. This technique, if properly integrated into the CA design, would fulfill the promise of building truly "conversational" agents. My findings indicate that language analysis can be used to automatically infer a community's perception of a community-facing CA. This opens up the potential of using language analysis to design CAs that can automatically identify the user's mental model of the CA, which allows the CAs to provide subtle hints in responses to guide the user in adjusting their mental model of the CA for a continuous and efficacious conversation.

In my study, even though JW is a question-answering(QA) CA designed to only fulfill students' basic informational needs, I was able to infer student perceptions through language features extracted from these simple QA dialogues. My findings resonate with prior work that also revealed the potential of using language analysis on question-answering conversational data between users and QA agents to infer conversation breakdowns [80]. I believe that in more sophisticated conversational settings where the human-CA interactions go beyond basic informational needs, and interactions that involve multimodal data (e.g., voice and visual communications), one can extract more nuanced descriptions of user perceptions about CAs. This would lead us to draw insights that can facilitate constructive and consistent human-CA dialogue.

I also note that student-JW interactions were situated in a much more controlled environment compared to many possible settings for human-CA interactions. For instance, the discussions in the online course forum are *supposed* to be thematically coherent about course work. Additionally, students are expected to self-present in a desirable and civil fashion—there are various online and offline norms and conventions that people tend to

follow in academic settings [312]. On the other hand, discussions on a general-purpose online community (e.g., Reddit), including those which are moderated, can not only have diverse and deviant discussions but can also include informal languages [313]. These kinds of data can add noise to automated language models, and it opens up more research opportunities to examine how language in general-purpose online communities reflect the individual and collective perception about a community-facing CA.

Besides helping CAs understand how they are perceived by the users during interactions, language can also potentially indicate user preferences about the CA in a particular context and thus inform future design of CAs. For example, in my regression analyses, linguistic measures such as sentiment and diversity reflect similar directionality (see Table 5.3) among the correlation between the three perception measures— I find a positive association between JW's perceived intelligence and likeability, but weak correlation between anthropomorphism and likeability. In particular, sentiment extracted from the student-JW conversation is significantly associated with both intelligence and likeability, yet not significantly associated with perceived anthropomorphism. It is thus worth considering whether human-likeness is a more important factor to consider comparing to an agent's intelligence demonstrated through providing informational support when designing virtual teaching assistants like JW. This finding also provides more evidence to the long-standing debate of whether CAs should be designed as humanlike as possible [266, 314], suggesting that user's preference of whether CAs should be humanlike is highly dependent on CA's role and use contexts.

5.5.2 Designing for Adaptive Community-Facing AI Agents

Prior work proposed seven social roles that community-facing CAs could serve within online human communities [270] yet how to quickly detect and measure people's perceptions and expectations of how the CA should behave when serving different social roles remained unexplored. This work opens up the opportunity to operationalize the desired social roles

of community-facing CAs in terms of specific dimensions of CA perceptions. For example, when CA serves as a social organizer to help community members build social connections, the community could expect the CA to behave more humanlike and more likeable instead of more intelligent. These expectations could potentially be identified and monitored through linguistic cues, as demonstrated by our work. This operationalization can help community-facing CAs quickly identify the community's expectations and produce behaviors that are better aligned with their perceived social roles within the community.

While prior research suggested community-facing CAs' shifting social roles over time within online communities [315, 270, 316], my examination of long-term changes in the student community's perception about JW provides empirical evidence on the specific variations in the community's perception of the agent. Our findings indicate that community-facing CA's perceived anthropomorphism and intelligence are more nuanced and fluid characteristics and thus require more frequent assessment for the CAs to adjust their behaviors within the community accordingly. JW's perceived likeability did not change significantly in our study, suggesting that designers could have more leeway in monitoring CA's perceived likeability. However, the reasoning behind JW's stable perceived likeability within the community requires further examination— it could be because long-term likeability perception is highly dependent on first-impression [282], or it could be a result of JW's stable performance over the semester due to its lack of learning ability.

One foreseeable challenge when designing adaptive community-facing CAs using linguistic cues to construct user perception of the CA is to distinguish the intention of each message— whether the user asked a genuine question or just trying to game the system; or whether the user's reply was intended for the CA or other community members. While people employ strategies such as changing appearances to manage their self-presentation in daily lives [30], people also manage their self-presentation through linguistic cues on public online platforms, depending on the perceived audience [317, 318, 319]. For community-facing CAs, every dyadic human-CA interaction is visible to other community members as

well. People thus might take advantage of this opportunity to not only gain support from the CA but also to modulate their responses to help manage their self-presentation within the community. For example, people might intentionally limit their emotional expression through language so that they don't appear "stupid" for thinking a CA could interpret the emotional elements in the language [312]; or people might purposefully reply with questions that can help them appear more humorous than to receive a correct answer from the CA. There are several occurrences of this in our study when students ask JW questions that are clearly out of scope for JW, such as *"What is the meaning of life?"* or *"What is your favorite character in Game of Thrones?"*.

5.6 Limitations and Future Work

This work has some limitations. The results might not be transferable when human-CA interaction takes place in private dyadic interaction contexts. This work investigates the feasibility of inferring student perception of a community-facing CA through linguistic features extracted from dyadic human-agent interaction on a public discussion forum. Student perception and interaction with the agent thus might be biased by other students' interactions with the agent on the public forum, which I point out as a unique challenge to design for community-facing CAs that carry out dyadic interactions within human communities. Future research aimed at designing adaptive CAs in dyadic interactions could replicate the current study in one-to-one human-CA interactions.

This work took a formative step towards understanding people's perception of a CA through linguistic features. The findings are correlational and we cannot make causal claims. Future work that accounts for unobserved confounds can lead to better insights into human-AI perceptions and interactions. We also recognize more qualitative or mixed-methods approaches are needed to gain deeper insights into people's reasoning and intention behind their linguistic behaviors when conversing with a CA. For example, in this study, students could be intentionally testing if JW learned anything from their previous questions by post-

ing the exact same questions from previous JW threads; or students might be frustrated by JW's learning ability and thus intentionally post difficult questions on the public thread — there is no way to evaluate this quantitatively, and future qualitative research could shed light on this issue.

To quantify student's perception of JW, I used a standardized measure taken from human-robot interaction that includes anthropomorphism, intelligence, and likeability [282]. However, the measurement I adopted does not suggest that these are, or should be, the standard dimensions of user perceptions of CA— in fact, prior research already suggested that there are different interpretations of how users build their mental models of CAs [96, 97]. I am, however, hopeful that language analysis can reveal the different dimensions of people's perceptions about CAs during interactions. Future research should replicate the current study using different measurements of the user's mental model about CA to provide more evidence on the potential of language analysis.

Finally, this work took place in a computer science class with computer science master students. These students perceptions of CAs might be more prone or less prone to changes than students who have less technical background, and these students might communicate with the CAs using different vocabularies (e.g., simpler words) than other students with non-technical background. Future work is required to understand the generalizability of these findings to a broader set of students with more diverse background and varying levels of technical skills.

5.7 Reflections & Takeaways

This chapter explores the first stage of the Mutual Theory of Mind framework for human-AI communication— ToM construction: AI's construction of human's interpretation of the AI. Through a longitudinal survey study while deploying an AI agent as a virtual teaching assistant in an online class, this study showed that students' collective perceptions of the AI agent evolve over time even when the AI does not have any learning capability

over time. I also found that linguistic characteristics such as readability and verbosity of the students' utterances to the AI can reflect their perceptions of the AI (e.g., AI's intelligence). This work established the feasibility to equip AI systems with ToM-like capability of constructing humans' perceptions of the AI through linguistic characteristics of humans' communication feedback.

While this chapter presents a preliminary study demonstrating the validity of extracting people's perceptions of AI through linguistic cues embedded in their utterances, future work should examine this technique further in more diverse human-AI communication contexts. Current study focused on the student community's perception of a single AI agent through postings on public discussion forum in a computer science class. Hence the findings might not be generalizable to more prevalent form of daily one-to-one human-AI communication, especially for users with varying levels of AI literacy who might communicate and perceive AI agents differently than students in the OMSCS program. In one-to-one human-AI communications, individual's perceptions of the AI agent might also change more rapidly due to frequent back-and-forth conversation turns between the individual and the AI agent. This presents an opportunity for future work to explore techniques that can enable AI agents' rapid calibration of people's perceptions of the AI agent at each conversation turn.

For AI system to better construct and calibrate people's perception throughout human-AI communication, it is critical to understand how people change their perceptions of the AI agent through the AI's communication feedback. In MToM in human-AI communication where both the human and the AI possess ToM-like capability, the human is able to recognize how they are interpreted by the AI through AI's communication feedback, which could cause a shift in people's perception of the AI. However, it is not yet clear how much or how often such perception shifts occur during human-AI communication where AI can communicate its interpretation of human characteristics. Understanding the frequency, magnitude, and possible factors that could prompt people's perception change after recog-

nizing AI's interpretation of them can provide design and technical implications on how AI systems should predict and respond to people's changing perceptions of the AI. The next chapter presents two studies that explored this problem in the ToM recognition stage by examining student's reactions and perception changes of the AI agent in the face of AI misinterpretation in AI-mediated social interaction.

CHAPTER 6

TOM RECOGNITION: HUMAN'S RECOGNITION OF AI'S INTERPRETATION OF THE HUMAN

This chapter examines the second stage of the MToM framework for human-AI communication in AI-mediated social interaction: *ToM Recognition: Human's Recognition of AI's Interpretation of the Human*.

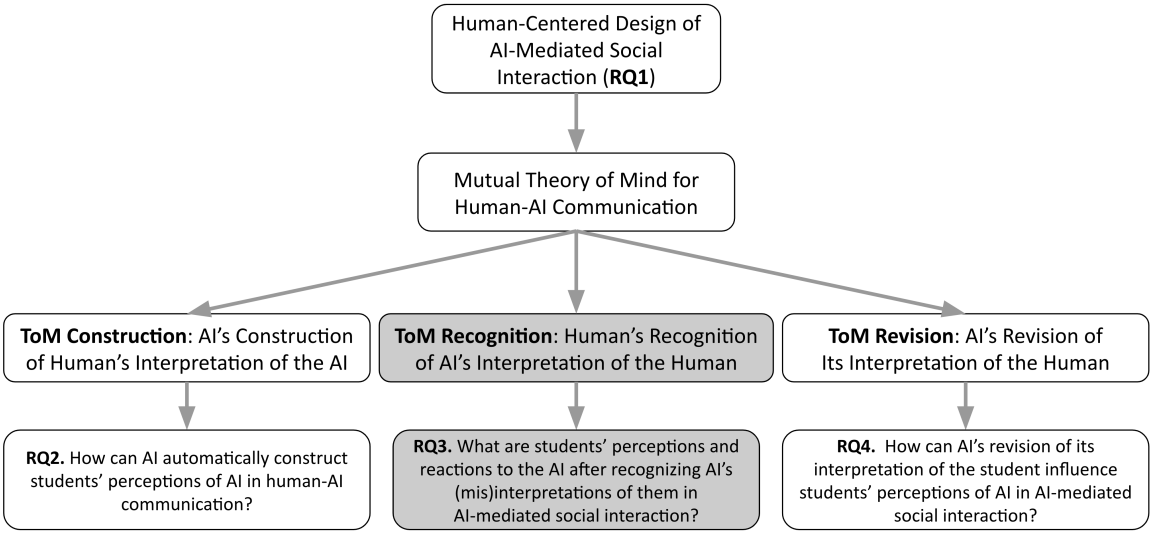


Figure 6.1: Chapter 6 explores ToM recognition: human's recognition of AI's interpretation of the human.

The previous chapter explored the ToM construction stage from the AI's perspective to examine how communication cues from the human's communication feedback can inform the AI's construction of human's interpretation of the AI. The ToM recognition stage, on the other hand, emphasizes on the human's perspective and describes the ToM recognition process of how AI's communication feedback can inform the human's recognition of the AI's interpretation of the human. In AI-mediated social interaction, humans are able to recognize AI's interpretation of their characteristics such as social preferences and needs through AI's social recommendations, as shown in Figure 1.2 in chapter 1. For example, if

the AI recommended a student with others who are interested in hiking, the student could infer that the AI has interpreted their hobby as hiking, and that the AI interpreted their social needs as connecting with others enjoying similar hobbies. While such personalized experiences provided students with much convenience to find social connections, students also expressed their concerns about being misinterpreted by the AI system in the studies conducted in chapter 4.

In AI-mediated social interaction, even AI systems with supposedly high accuracy in interpreting humans can make mistakes and misinterpret people's characteristics. Such AI misinterpretations, or misrepresentations, in AI-mediated social interaction can not only have negative consequences on people's perceptions of AI [22], but also cause concerns and potential reputational harms for students, as suggested by my studies described in chapter 4. Understanding people's reactions and perceptions of the AI after encountering AI misrepresentations could offer valuable insights into whether and how people changed their intuitions, beliefs, and reactions of AI in the face of AI misrepresentations, providing critical implications for the human-centered design and development of mitigation and repair strategies to minimize potential harms when such AI systems inevitably err. This chapter examines the following research question:

RQ3. What are students' perceptions and reactions to the AI after recognizing AI's (mis)interpretations of them in AI-mediated social interaction?

In this chapter, I examined this question by contextualizing it in one type of AI-mediated social interaction— AI-facilitated team matching. There has been an increasing use of hyper-personalized AI systems that can recognize students' personality traits to facilitate school project team formations in higher education [320, 321, 322, 323]. However, students' reactions and perceptions of AI misrepresentation of their most intimate personality traits have not been explored. Through semi-structured interviews and a large online survey experiment, I found that people's existing and newly acquired AI knowledge plays a critical role in shaping their perceptions and reactions after encountering AI misrepresenta-

tions. Findings from this chapter provides a descriptive account of how people navigate AI's misinterpretations through people's evolving AI knowledge and provided implications for designing and developing responsible mitigation strategies that consider people's evolving AI knowledge to reduce potential harms when AI fails to capture accurate interpretations of people's characteristics.

6.1 Introduction

Recently, a plethora of hyper-personalized AI systems that can profile users' characteristics and traits have been deployed in people's daily lives, with the ultimate goal of providing personalized shopping, music, and social media recommendations. As these systems become more advanced in profiling people's most personal and complex traits such as personalities and emotions [33, 99, 32], they sometimes give people the illusion that "machines can read our minds" [46]. This illusion has led to various—rather concerning—reactions and perceptions of AI with people attributing AI with beyond-human expertise at reading people's emotions and personalities [47, 45]. However, people's perceptions and reactions of AI when this illusion is broken in the face of *AI misrepresentations* have not yet been explored.

AI misrepresentations is one type of AI fallibilities when AI misinterprets people's most intimate and complex traits like personality and emotions, aspects where people possess the most self-awareness. Prior work has suggested that AI mistakes on human-AI tasks could erode people's trust and social perceptions (e.g., anthropomorphism, intelligence, likeability) of the AI [135, 324]. However, when faced with AI misrepresentations of people's most intimate traits, people may dismiss it based on their self-awareness, resulting in lack of adherence and trust in AI; or people might exhibit unwavering trust in AI, allowing it to persuade them into accepting false information about themselves.

To understand people's reactions and perceptions of the AI after encountering personality misrepresentations by AI, I seek to explore three research questions:

- RQ3.1. What perceptions and reactions do students have about the AI after encountering AI misrepresentations of their personality traits in AI-facilitated team matching?
- RQ3.2. How do students change their perceptions of the AI after encountering AI misrepresentations of their personality traits in AI-facilitated team matching?
- RQ3.3. What factors contribute to students' perception changes after encountering AI misrepresentations of their personality traits in AI-facilitated team matching?

To answer these research questions, I conducted semi-structured interviews with twenty college students (Study 1) and a large survey experiment (Study 2) with 198 students on the Prolific platform. In both studies, I took a Wizard-of-Oz approach to fabricate intentionally inaccurate/accurate personality inferences based on participants' personality ground truth. I showed participants in both studies their "AI-generated personality inferences" to elicit their perceptions and reactions of AI misrepresentations. I found that people's existing and newly acquired AI knowledge plays a critical role in shaping their perceptions and reactions after encountering AI misrepresentations. Specifically, I pinpointed three rationales that people adopted through knowledge acquired from AI (mis)representations: AI works like a machine, human, and/or magic. These rationales are highly connected to their reactions of over-trusting, rationalizing, and forgiving of AI misrepresentations. I also found that people's existing AI knowledge, i.e., AI literacy, significantly moderate the level of changes in people's overall trust after encountering AI misrepresentations.

6.2 Study Overview

To examine people's reactions and perceptions of AI after encountering AI misrepresentations, I conducted two studies using a mixed-methods approach. The first study (Study 1) focuses on qualitatively exploring students' perceptions and reactions to AI after encountering AI misrepresentations, and the second study (Study 2) focuses on quantitatively

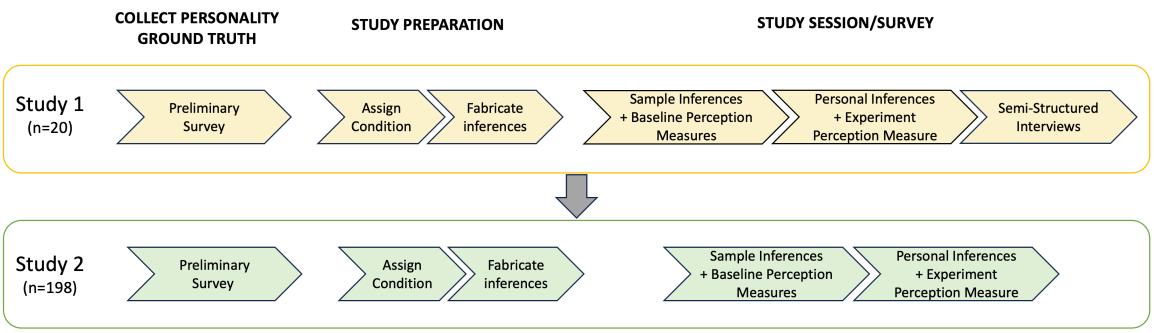


Figure 6.2: Study flow diagram that shows the procedures of Study 1 and Study 2. Study 2 occurred after Study 1 was concluded. All personal inferences shown to participants were either accurate or inaccurate based on the condition assigned to the participants.

examining the factors, specifically AI literacy, contributed to the variations in students' perceptions of the AI after encountering AI misrepresentations. Given that AI output is often non-deterministic and thus difficult to control and manipulate, I took a Wizard-of-Oz approach to fabricate and control the accuracy of the AI inferences— in both studies, human researchers fabricated all the AI inferences based on participants' personality ground truth, collected in the preliminary survey in each study. To prevent raising participants' suspicions when shown fabricated AI inferences, they were not presented with the inferences until at least a week after they completed the preliminary survey.

To compare and contrast participants' reactions and perceptions of AI misrepresentations, I divided all participants into two conditions in both studies: participants in the accurate condition received accurate "AI inferences," and the inaccurate condition received inaccurate "AI inferences." In both studies, participants were asked to evaluate AI-generated inferences based on students' self-introduction paragraphs, which will be used by an AI agent named SAMI (stands for "Social Agent Mediated Interaction") to match them with potential teammates for a school project.

In Study 1, I conducted user study sessions to understand students' reactions and perceptions (RQ3.1, RQ3.2) of the AI after encountering AI (mis)representations by showing them different SAMI inferences, including inferences about themselves. Based on our observations in Study 1, I revised some of the measurements and replicated Study 1 as a

survey experiment without the semi-structured interview part. I conducted Study 2 on the Prolific crowdsourcing platform to obtain a larger sample to examine changes in students' perceptions (RQ3.2) and factors that contributed to those changes (RQ3.3). Both studies were approved by the IRB at Georgia Institute of Technology. Before Study 1, I conducted a pilot study with 14 people to test out the study procedure and the SAMI inference fabrication process.

6.3 Study 1: Understanding Students' Perceptions and Reactions to AI Misrepresentation

In this section, I first describe the method I used in Study 1, including recruitment, study setup, and study procedure. I then talk about my data analysis process. Finally, I present my findings on students' perceptions and reactions to SAMI misrepresentations.

6.3.1 Study 1 Method

Study 1 Participant and Recruitment

I conducted remote 60-minute user study sessions with 20 current students, all recruited from a large public U.S. technology institute. I recruited the participants by posting the recruitment message on the institute's Reddit and by broadcasting the recruitment email to different departments and programs, especially non-STEM programs, within the institute to increase sample diversity. My recruitment message invited students to evaluate an AI agent SAMI that could perform team matching by drawing inferences from students' self-introduction. Students signed up for the study by filling out a preliminary survey.

In the preliminary survey (see Appendix section C.3), students wrote a paragraph of self-introduction as a free-flowing essay to introduce themselves to SAMI. They then completed a 44-item Big Five personality survey measurement to provide ground truth of their personality. Students were then asked about level of technology proficiency ("Beginner", "Intermediate", or "Expert."), general attitude towards AI technology on a scale of "1-Very

negative" to "5-Very positive", as well as their demographic information, level of study, major, academic or professional background.

I received 110 valid preliminary survey responses. I sent out follow-up invitations in batches to balance sample diversity and ended up with 20 participants in total. I then evenly assigned these participants to either the inaccurate or accurate condition (10 participants in each condition) while balancing the sample diversity in each condition. The median age of the participants was 20 years old. There were 15 undergraduate students and 5 master's students in my study. Their gender breakdown is as follows: women (n=10), men (n=7), non-binary (n=2), one participant did not report their gender. Participants came from a variety of majors, with mostly intermediate tech proficiency, and varying general attitudes toward AI. Table Table 6.1 shows the information for all the participants in Study 1.

Fabricating SAMI Inferences

I took a Wizard of Oz approach and let human researchers fabricate SAMI's personality inferences to control for inference accuracy. The fabrication of SAMI inferences was based on (1) participants' ratings of each statement on the Big Five personality test, and (2) the condition (either accurate or inaccurate) the participant was assigned to. Participants in the accurate condition would receive accurate inferences, which would consist of all the statements they rated as "4-agree" or "5-strongly agree" in the personality test, or the reverse of the statements they rated as "1-strongly disagree" or "2-disagree" in the personality test; Participants in the inaccurate condition would receive inaccurate inferences, which would consist of all the statements they rated as "1-strongly disagree" or "2-disagree" in the personality test, or the reverse of statements that they rated as "4-agree" or "5-strongly agree" in the personality test. I illustrated this fabrication process in Figure 6.3 below.

When selecting statements from the personality test, I picked the three dimensions with the most extreme scores, while excluding dimensions with a neutral score of three out of five. I also restricted the length of SAMI's inferences for each participant to about 10

Table 6.1: Study 1 participant information. In the Gender column, "W" stands for "Woman", "M" stands for "Man", "NB" stands for "Non-Binary." In the Level of Study column, "UG" stands for "Undergraduate." In the Major column, "Eng." stands for "Engineering", "Comp." stands for "Computational." In the Tech Proficiency column, participants self-reported their technology proficiency as "Beginner", "Intermediate" or "Expert." In the Attitudes Toward AI column, participants self-reported their attitudes toward AI on a scale 1-5: 1-Very Negative, 2-Neutral to Negative, 3-Neutral, 4-Neutral to Positive, 5-Very Positive.

Condition	**ID**	**Age**	**Gender**	**Study Level**	**Major**	**Tech Proficiency**	**Attitudes Toward AI (1-5)**
Accurate Condition (n=10)	P14	22	W	UG	Psychology	Intermediate	4
	P19	18	W	UG	Psychology	Intermediate	2
	P22	28	W	Master	Digital Media	Intermediate	3
	P26	21	M	Master	Digital Media	Intermediate	3
	P29	19	W	UG	Neuroscience	Intermediate	2
	P30	31	W	Master	HCI	Expert	4
	P33	19	M	UG	Chemical Eng.	Intermediate	4
	P34	19	NB	UG	Comp. Media	Intermediate	5
	P37	20	M	UG	CS	Expert	5
	P40	19	M	UG	Computer Eng.	Intermediate	5
Inaccurate Condition (n=10)	P16	19	W	UG	Psychology	Intermediate	4
	P17	21	W	UG	Psychology	Intermediate	3
	P21	19	NB	UG	Psychology	Expert	4
	P23	20	W	UG	Psychology	Intermediate	2
	P24	23	W	Master	HCI	Intermediate	4
	P28	20	W	UG	Business Admin	Intermediate	4
	P35	22	M	Master	CS	Expert	4
	P36	22	M	UG	Industrial Eng.	Expert	3
	P39	21	No Report	UG	Biology	Intermediate	4
	P41	18	M	UG	Computer Eng.	Expert	4

statements to remove the potential effect of inference length on participants' perception of SAMI. In the Big Five personality test, some statements were inherently positive and some statements were inherently negative. To remove the potential effect of the sentiment of SAMI inferences on students' perceptions and reactions to SAMI, I composed each SAMI inference with about 40% negative inference and 60% positive inference. I composed an inference fabrication guideline to document these rules and procedures (see Appendix section C.1) and closely followed the guideline when fabricating SAMI inferences.

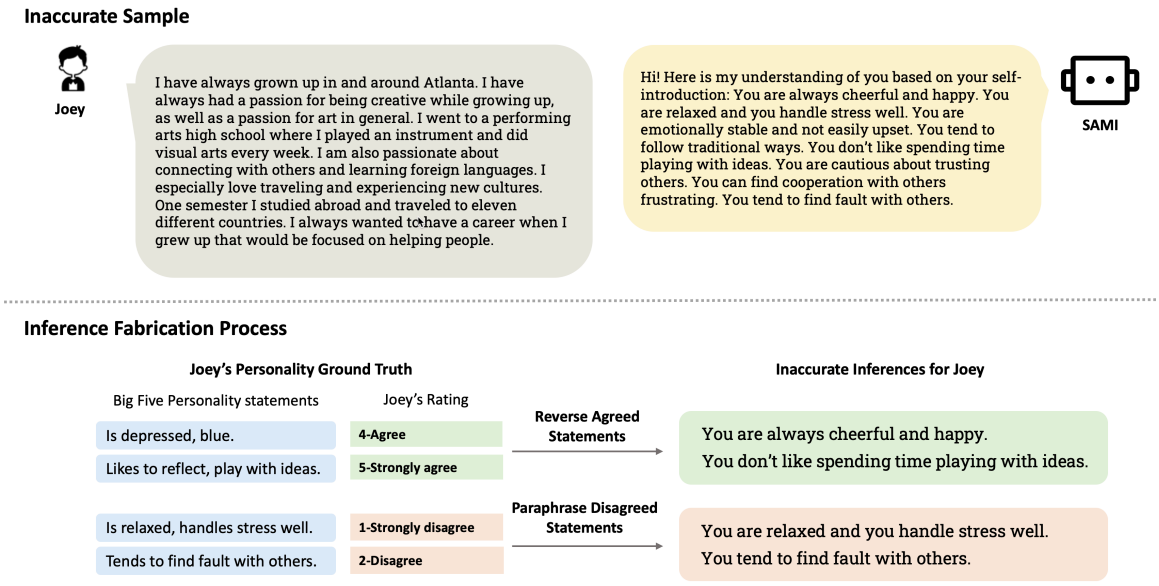


Figure 6.3: This figure shows the sample and my inference fabrication process for the sample student. The top half of this figure shows one of the samples I showed to the participants that is inaccurate. The bottom half of this figure shows how I utilized participants' personality ground truth filled out in the preliminary survey to fabricate inferences for them based on the condition they were assigned.

I began each session by emphasizing to the participants that they should share their honest opinions, both positive and negative. Each study session had two parts. In the first part, participant was shown two samples of self-introductions and SAMI's inferences, then asked to fill out perception measurements (see Appendix section C.4) to record their baseline perceptions of SAMI in terms of trust [325] and social perceptions (likeability, perceived intelligence, anthropomorphism) [282]. Participant was then shown their own self-introduction and SAMI's inferences about them, then asked to fill out the same measurements. The goal was to record students' perceptions of SAMI immediately after each SAMI inference was presented to them. The two samples (see Appendix section C.2) were written by real students from the pilot study. SAMI's inferences about the sample students were fabricated in the same way as described in Appendix section section C.1. SAMI's inferences for one of the samples were generated to be inaccurate, and the other one accurate. The samples were shown in random orders to remove potential order effects.

The second part of the session was a semi-structured interview to go through partici-

pants' reactions and perceptions of SAMI after encountering AI (mis)representations. During the interviews, I asked participants to walk me through their reactions to each of the SAMI inference, what they thought about SAMI after seeing each inference, and how they thought SAMI extracted the inferences. Throughout the interview, I used the perception measurements as a probe to provide basis for them to elaborate on their perceptions of SAMI after seeing each inference. I then debriefed the participants about the fabrication process of SAMI's inferences. I addressed any comments and questions that participants had and then provided compensation of USD $25 gift card. All the study sessions were conducted remotely on Zoom and all lasted around 60 minutes. The session and interview protocol is attached in Appendix section C.2.

6.3.2 Study 1 Data Analysis

All sessions were video recorded and transcribed for data analysis. I adopted Braun and Clarke (2006)'s Reflexive Thematic Analysis (RTA) approach [326, 327] in my analysis. RTA encourages researchers to embrace their stance and assumptions during analysis. This differs from other qualitative data analysis approaches such as grounded theory and codebook approach that value objectivity and removing researcher bias, which does not allow enough flexibility for researchers to actively participate in the data analysis process.

Two researchers participated in the analysis to collaboratively discuss and iterate on themes that emerged from the data using RTA approach. We followed the analysis process outlined in Braun and Clarke (2006). We generated our initial codes in two rounds by first dividing the 20 transcripts among the two researchers for independent coding, then swapped the transcripts for a second round of independent coding. After the initial codes were generated, the two researchers frequently met to discuss, review, and search for themes. The first round of analysis generated 16 domain categories (e.g., folk theories of SAMI) and 163 codes (e.g., believing SAMI's misrepresentation). By comparing our second round codes with the first round codes and domain categories, we distilled five themes

(e.g., tendency to overtrust AI) and 27 codes (e.g., attributing human inference-making process to SAMI). We continued to review and search for larger themes and ended up distilling two bigger themes to understand students' perceptions and reactions to SAMI after encountering AI misrepresentations, which I present in detail in the next section.

6.3.3 Study 1 Findings: Interpreting and Reacting to AI after Encountering AI (Mis)representation

I first confirmed that our manipulation of SAMI inference accuracy was largely effective. I compared students' baseline accuracy rating of SAMI's inferences after seeing the samples and their experiment accuracy rating of SAMI's inferences after seeing their own inferences. I found that in the accurate condition, the accuracy rating largely increased from baseline rating to experiment rating, with a median increase of 1 out of 5; in the inaccurate group, the accuracy rating largely decreased from the baseline rating to the experiment rating, with a median decrease of 0.5 out of 5. However, it is worth mentioning that some participants' accuracy ratings did not change between baseline and experiment (n=1 in the accurate condition, n=3 in the inaccurate condition).

Next, I present the findings from Study 1 to understand students' reactions and perceptions of SAMI after encountering AI (mis)representations. I first describe three rationales that participants adopted through knowledge acquired from SAMI (mis)representations to interpret how SAMI worked: SAMI works like a machine, a human, and/or magic. I then describe participants' reactions of over-trusting, rationalizing, and forgiving of SAMI misrepresentations, highlighting that these reactions were highly connected to the rationales that we pinpointed.

Interpreting SAMI: Machine? Human? Or Magic?

Based on the interviews with the participants, I found that participants acquired new knowledge from SAMI (mis)representations. Such newly acquired knowledge prompted participants to adopt different rationales to interpret how SAMI worked: SAMI works like

a machine, human, and/or magic. These rationales could co-exist at any given time, yet are bounded by participants' existing AI knowledge, tech proficiency, and how much they could make sense of SAMI's specific inferences.

"SAMI works like a machine." The first rationale participants held was that SAMI followed the typical "input-processing-output" machine working mechanism. About a quarter of the participants in the study, half of whom self-rated as "Expert" in terms of their tech proficiency, could clearly explain how SAMI came up with specific inferences by specifying their perceived SAMI knowledge base, training, input processing, and inference generation process. P23 described her speculation of how SAMI processed the self-introduction input to generate personality inferences: *"From a personality standpoint, I'd imagine that it's sectioning out facts and being like, people who do this are commonly [like] that. And then kind of filtering through several things that point to this person being generally trusting. So in this sample, this person want to travel, want to visit the seven wonders of the world, maybe all of that means they are generally trusting."*

"SAMI works like a human." I also observed the "SAMI works like a human" rationale surfacing among some participants, all of whom self-reported their tech proficiency as "Intermediate." These participants believed that SAMI drew personality inferences in similar ways as how humans would do it. For instance, P14 described how SAMI worked like humans when coming up with inferences: *"I think they are probably like us, where [certain things] kind of tick SAMI off. Like maybe people who love to travel can be a little bit more easily distracted, or maybe a little bit more careless, just because they have to be more go with the flow. "* When P16 also described her rationale of how SAMI came up with inferences like humans would, she elaborated: *"So it's funny, because I know [humans and AI] are not the same at all. But I think like, who's making the AI technology— people! So maybe the people try to make it think like how we think, even though it's not one and the same. Maybe whoever coded it tried to make it like human perception. "*

"SAMI works like magic." When participants couldn't make sense of SAMI's unexpected inferences from the given self-introduction paragraph, they would sometimes resort to the belief that SAMI was a powerful, knowledgeable, "big black box of magic." This rationale was mentioned by participants with varying levels of tech proficiency. Participants who adopted the magic rationale often could not clearly articulate how SAMI came up with the specific inferences, but instead loaded their answers with vague terms to emphasize the massive amount of training data SAMI might have access to or the "magical power" of AI to identify patterns that humans cannot see. For example, P36 who self-rated as an "Expert" in tech proficiency explained why he thought some of SAMI's inaccurate inferences about him were true: *"I was under the impression that it somehow drew inferences... since I didn't see where it got that from, I thought it was drawing inferences from different sources of information, and somehow using a pattern, like using a neural net or something. So I was like 'Ok, sure maybe I can see that.' But if a person just told me that and I just don't see where they got it from, then I'd just think the person is being inaccurate. (P36)"*

Reacting to AI Misrepresentation: Over-trusting, Rationalizing, and Forgiving

After being shown SAMI's inaccurate personality inferences about them, participants displayed a range of reactions: some participants believed there was some truth to SAMI's misrepresentation; some participants rationalized it and blamed themselves instead; some participants were forgiving of SAMI's mistakes. These reactions to SAMI misrepresentations seemed to be connected to the rationale participants adopted after acquiring new knowledge from SAMI misrepresentations about them.

Over-trusting SAMI Misrepresentations. To my surprise, I found that most participants in the inaccurate condition often found some truth to SAMI's inaccurate inferences about them. Given that I made every SAMI inference to be intentionally inaccurate and the complete opposite of participants' personality ground truth, I expected most participants in the inaccurate condition to rate SAMI's inferences about them as "1-Not accurate at all." How-

ever, only two out of 10 participants did so— rest of the participants displayed varying levels of trust in SAMI's inaccurate inferences about them, ranging from two to "3-Somewhat accurate"

These participants often fell into the trap of the Barnum effect [328], a cognitive bias phenomenon that people tend to believe in personality descriptions of them as customized to them, when in fact, these descriptions are often vague and general. When presented with SAMI's inferences about them, participants had a tendency to look for evidence in their daily lives to support SAMI's claims. In addition, believing in the authority of the evaluator could make the Barnum effect stronger [328]. I found that participants in the inaccurate condition displayed a tendency to perceive SAMI as an authority figure that was "smarter and more powerful" than them, with opaque working mechanism.

This echoed with the "SAMI works like magic" rationale that they don't know how AI works, but it just should work. This rationale was even present when participants were confident that SAMI's inferences were very inaccurate. For example, P28 rated SAMI's inferences as very inaccurate, yet still took a huge sigh of relief when we told her SAMI's inferences about here were fabricated to be inaccurate: *"I feel better now because I was worried. I was worried that, because I don't see myself from the outside so I was worried that SAMI was real and kind of accurate, and this is how people see me. "* She further explained about her reasoning: *"I guess partly because it's an AI. Because AI is a lot smarter than me. And I don't consider myself to be a computer science expert. AI is so unfamiliar to me that I don't really know what is the line between true and false. I don't know when to trust it or not to. I think with SAMI, I was erring on the side of being not trusting but then deep inside, I was like, 'this is an AI, it's really smart, it should be able to know.' I think that's why I felt like I could trust it. After I saw the two samples and then when it came to mine, I started to feel like something was wrong. Uhm, but then deep down I was like, 'ok, well, maybe it is right."'*

Rationalizing SAMI Misrepresentations. Besides participants' tendency to over-trust AI

misrepresentations, I also found that participants had a tendency to rationalize AI mistakes, or "find excuses" for SAMI's inaccurate inferences. When participants noticed that SAMI's inferences might not be accurate for either the students in the sample or themselves, some participants would "justify" SAMI's inaccuracies, citing it could be a result of the nature and quality of the self-introductions. This echoed with the "SAMI works like a machine" rationale that participants believed "unqualified" input would hurt the machine's output. For instance, when P29 spotted SAMI's inaccurate inferences in one of the samples, she said, *"I feel like the person was fairly broad in what they wrote. And so that also could have influenced SAMI's response. Like it didn't necessarily go too in-depth, [...] It also went from learning languages, study abroad, and then the career of helping people. That could have thrown SAMI off a little bit as well, seeing as it like, it bounced around a little bit more. "* P36 also said that SAMI might have made inaccurate inferences about him due to him not mentioning certain things in his self-introduction: *"'You can be somewhat careless.' I would say that is like very inaccurate. I'm a very meticulous person. But I guess I never really touched on that anywhere in my response..."*

Forgiving SAMI Misrepresentations. Many participants were forgiving of SAMI's misrepresentations of either themselves or students in the samples. This echoed the "SAMI works like a human" rationale in that this reaction is analogous to how they would react to human mistakes. After encountering SAMI misrepresentations, participants attributed good intentions and efforts to SAMI. Following this line of reasoning, participants believed that just like human mistakes, SAMI's mistakes could be forgiven considering its good intentions and efforts. This reaction was presented in more than half of the participants in the inaccurate condition. After reviewing SAMI's misrepresentations of themselves, they believed that SAMI was "trying" and well-intentioned, but just not as capable as they expected. Some participants in the accurate condition, noticing that SAMI did not make perfect inferences in the samples, were still hopeful about SAMI despite its mistakes. P29 said, *"[I'm thinking of] a little kid where they want to help out and be like 'oh my gosh,*

look what I did' and presented it to you and be like 'Look, I'm trying to be helpful.' But it's not always the most helpful or accurate in this situation. I was thinking that [SAMI] is trying to help, it is generating those responses, but at the same time, it just might not be the most accurate."

6.4 Study 2: Examining the Changes in Students' Perception of AI after Encountering AI Misrepresentations

To explore the changes in students' perceptions of AI (RQ3.2 and RQ3.3) after encountering AI misrepresentations, I conducted a survey experiment on Prolific and solicited a larger sample to quantify the changes in student perceptions as well as contributing factors to the changes. Similar to Study 1, I first deployed a preliminary survey to collect information for fabricating students' personality inferences. Then after a week, I followed up with an experiment survey to show participants SAMI's inferences and measured their perception changes of SAMI in terms of overall trust, anthropomorphism, perceived intelligence and likeability. Based on my observations from Study 1, I added a general AI literacy measurement in the preliminary survey to examine AI literacy as a potential factor contributing to the changes in students' perceptions of AI after encountering AI misrepresentation.

6.4.1 Study 2 Study Design

I deployed the preliminary survey (see Appendix section C.8) on Prolific and recruited current students above age 18 of all study levels in the United States. Participants were told that the goal of this study was to understand students' perception of a team-matching AI agent. In the preliminary survey, participants provided their self-introduction paragraph, responses to the personality test and an AI literacy questionnaire, as well as their demographic, background, and team project experience information. Participants who completed the preliminary survey were compensated with $3 USD. The median completion time was 11 min and 41 seconds.

250 Prolific participants filled out our preliminary survey on Qualtrics. I removed seven participants whose self-introductions were suspected to be generated by generative AI tools such as ChatGPT. Given our emphasis on AI misrepresentation in Study 2, I assigned about 70% of the participants to the inaccurate condition and about 30% of the participants to the accurate condition. I then wrote a Python script to fabricate SAMI's inferences for the 243 remaining participants following the same fabrication procedure and guidelines in Study 1. By following the 60% positive and 40% negative inference ratio rule, I removed 21 participants since their fabricated inferences were either too positive or too negative based on their personality ground truth. This resulted in 222 participants remaining (n=157 in inaccurate condition, n=65 in accurate condition), all of whom were invited to participate in the experiment survey.

The experiment survey (see Appendix section C.9) followed the Study 1 procedure by showing the same two samples and prompted participants to answer their perceived accuracy of SAMI inferences, as well as filling out measurements of their trust and social perceptions of SAMI. Participants then retrieved SAMI's inferences about them by entering their Prolific ID on a website that we created (see Appendix section C.7), and then filled out the same perception measurements. Finally, a debriefing form informed participants about the real purpose of the study and how SAMI's inferences were fabricated. Participants were compensated with $3 USD for completing the experiment survey, and an extra $4 USD bonus for completing both surveys. I received 211 responses (n=151 for inaccurate condition, n=61 for accurate condition) for the experiment survey. The median completion time was 12 min and 43 seconds. A description of the perception measures and general AI literacy measures used in Study 2 can be found in Appendix section C.5.

After I concluded the data collection, I removed 13 participants' data due to extremely fast completion time (less than six minutes) and obvious contradictions in their responses. I ended up with 198 participants' data for the experiment survey, with 57 participants in the accurate condition and 141 participants in the inaccurate condition.

6.4.2 Study 2 Participant Summary

The final participant pool (n=198) has an average age of 31.3±11.12, ranging between 18-74 years old. 42.9% were women, 53.5% were men. Most students were at the undergraduate level (n=146, 73.7%). 46.5% students in non-STEM major, 52% in STEM major. Participants were relatively familiar with AI, with an average of 14.3±4.50 out of 25 on overall AI literacy. Participants were generally experienced in team projects at school, having participated in an average of 12 team projects (median=5, SD=25.8, range=0–300). Participants held a relatively positive attitude about their team project experience, reflected in an average rating of 3.7 ± 0.88 out of a five-point Likert scale. I provided more details about the participants' demographic and personality in Appendix section C.6.

6.4.3 Study 2 Data Analysis

I first calculated the changes in students' perceptions of SAMI (overall trust, anthropomorphism, perceived intelligence, likeability) by taking the difference between students' baseline perceptions (B_p) and experiment perceptions (E_p): $\Delta = E_p - B_p$. I then coded condition as 0 (accurate condition) and 1 (inaccurate condition) during analysis.

To understand the effect of AI misrepresentation on students' perception changes of SAMI (RQ3.2), I performed four sets of linear regressions with the Δ of each perception construct as the outcome and the participants' condition as the independent variable. I controlled for age, gender, study level, major (STEM or non-STEM), number of projects, project experience rating, and five personality dimensions. Equation 6.1 describes our linear regression models, whereΔ $\mathcal{P}$ refers to changes in overall trust, anthropomorphism, intelligence, and likeability.

$$\Delta\mathcal{P} \sim Condition + Age + Gender + StudyLevel + Major + ProjectCount + ProjectExperience + Extroversion + Neuroticism + Agreeableness + Openness + Conscientiousness \quad (6.1)$$

To understand the moderating effect of AI literacy (RQ3.3), I again performed four sets of linear regression models, but this time incorporating AI literacy. Specifically, I performed two versions of linear regression for each model: one that included AI literacy (Equation 6.2), and another included AI literacy plus an interaction effect between AI literacy and condition (Equation 6.3). I included the only significant covariate from prior models— the Openness personality dimension. The outcome $\Delta\mathcal{P}$ is the changes in overall trust, anthropomorphism, intelligence, and likeability.

$$\Delta\mathcal{P} \sim Condition + AILiteracy + Openness \quad (6.2)$$

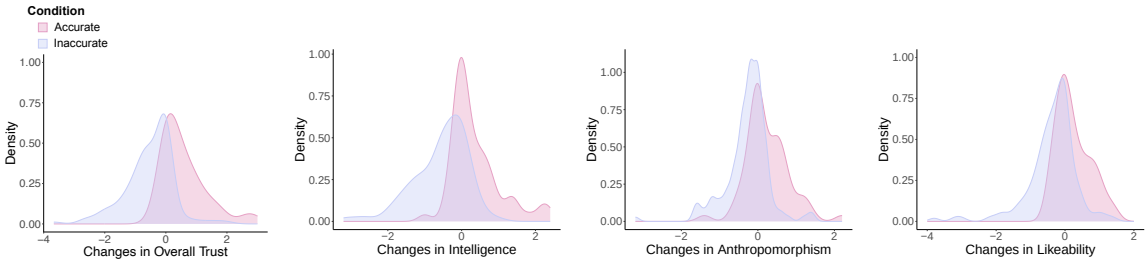
$$\Delta\mathcal{P} \sim Condition + AILiteracy + Condition * AILiteracy + Openness \quad (6.3)$$

6.4.4 Study 2 Findings

The results show that encountering AI misrepresentations had a significant effect on changes in students' overall trust, perceived intelligence, anthropomorphism, and likeability of SAMI. I also found that AI literacy could moderate the effect of AI misrepresentations on changes in students' overall trust in SAMI, but not their social perceptions of SAMI. I present the detailed findings below.

Examining the Effect of AI Misrepresentation on Changes in Students' Perceptions of AI

To understand the effect of AI misrepresentations on students' perceptions of SAMI, I first look at changes in their perceptions before and after encountering AI misrepresen-



(a) (b) (c) (d)

Figure 6.4: Density plots visualizing the participant distribution of changes in overall trust, intelligence, anthropomorphism, and likeability in the accurate and inaccurate conditions.

tations. Students' baseline perceptions of SAMI showed that participants in both conditions had similar initial perceptions in terms of overall trust (accurate condition median=2.33, SD = 1.04; inaccurate condition median=2, SD = 1.02), intelligence (accurate condition median=3.4, SD=1.09; inaccurate condition median=3, SD=0.95), anthropomorphism (accurate condition median=2.2, SD=1; inaccurate condition median=2.2, SD=0.83), and likeability(accurate condition median=3.2, SD=1.01; inaccurate condition median=3, SD=0.79).

I then looked at the changes in students' perceptions after I introduced SAMI's inferences in the two conditions. I plotted out the changes in each perception outcome in density plots as shown in Figure 6.4. I noticed that many participants in the inaccurate condition decreased their perceptions of SAMI after encountering AI misrepresentations compared to participants in the accurate condition. Therefore, encountering AI misrepresentations seemed to have a negative effect on changes in students' perceptions of SAMI.

To further examine the effect of encountering AI misrepresentation on students' perceptions of SAMI, I built four linear regression models with changes in each perception as the outcome. Our models (based on Equation 6.1) show that after controlling for demographics, team project numbers and experiences, as well as the personality dimensions, encountering AI misrepresentation had a significant effect on changes in students' perceptions of SAMI in terms of overall trust, perceived intelligence, anthropomorphism, and likeability (Table 6.2). Specifically, participants who encountered AI misrepresentations

Table 6.2: Results of our regression models(Equation 6.1) show that participants in the inaccurate condition had a significant decline in overall trust, perceived intelligence, anthropomorphism, and likeability. The only significant covariate, the Openness personality dimension, is reported in the table. *** p<0.001 ** p<0.01 * p<0.05 . p<0.1

	Overall Trust			**Intelligence**		
	Est.		S.E	Est.		S.E
(Intercept)	0.84		0.782	0.64		0.728
Openness	-0.08		0.100	-0.17	.	0.093
Condition (Inaccurate)	-1.14	***	0.133	-0.92	***	0.124
	Adj. R^2=0.283***			Adj. R^2=0.234***		
	Anthropomorphism			**Likeability**		
	Est.		S.E	Est.		S.E
(Intercept)	-0.06		0.578	1.39	*	0.664
Openness	-0.09		0.074	-0.18	*	0.085
Condition (Inaccurate)	-0.50	***	0.098	-0.58	***	0.113
	Adj. R^2=0.111***			Adj. R^2=0.139***		

reported significantly lower overall trust in SAMI (Est.=-1.14, $p < 0.001$) in comparison to those in the accurate condition. I also found that participants who encountered AI misrepresentations reported significantly lower perceived intelligence (Est.=-0.92, $p < 0.001$), anthropomorphism (Est.=-0.50, $p < 0.001$), and likeability (Est.=-0.58, $p < 0.001$) compared to the participants in the accurate condition. This suggested that after people encountered AI misrepresentation, they were more likely to view the AI as less trustful, less intelligent, less humanlike, and less likable.

AI Literacy as a Moderator on the Effect of AI Misrepresentation on Students' Perceptions of AI

I then looked at the possible moderating effect of AI literacy on the effect of AI misrepresentation on the changes in students' perceptions of AI. I performed two versions of each of the four models to examine AI literacy's moderating effect (Table 6.3). As shown in Table 6.3, I performed two versions of linear regression (base models, and the base+interaction models) for each of the four models to examine AI literacy as a moderator

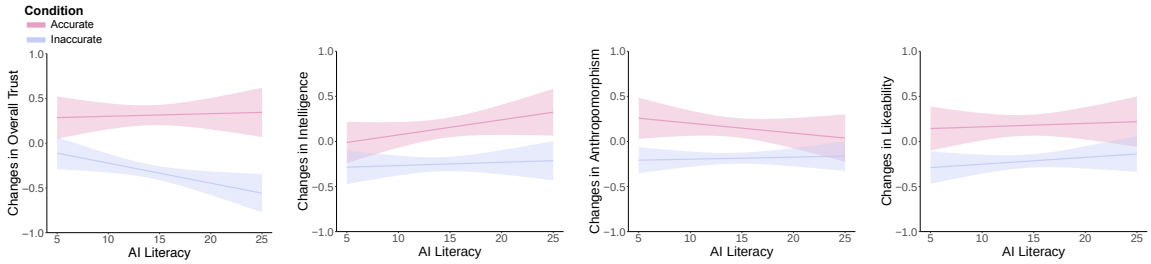
on changes in students' perceptions after encountering AI misrepresentations.

The base models that include an effect for AI literacy (Equation 6.2) (Models 1a., 2a., 3a., 4a., in Table 6.3) show that AI literacy alone does not have a direct relationship with the outcomes; however, the models that include an interaction effect between AI literacy and condition (Equation 6.3) (Models 1b., 2b., 3b., 4b., in Table 6.3) show that there is a significant interaction between AI literacy and condition on changes in overall trust (Est. = -0.06, $p < 0.05$). However, this interaction is not significant on changes in intelligence, anthropomorphism, and likeability. Detailed model results can be found in Table 6.3. This suggests that students' AI literacy could have an effect on students' overall trust of AI after encountering AI misrepresentations; however, students' AI literacy does not have an effect on students' perceived intelligence, anthropomorphism, likeability of the AI after encountering AI misrepresentation.

Table 6.3: Results of the regression models with AI literacy in base models (Equation 6.2) and base + interaction models (Equation 6.3). Results suggested a significant interaction effect between condition and AI literacy in changes in overall trust after encountering AI misrepresentations. However, a significant interaction effect is not found in changes in intelligence, anthropomorphism, and likeability *** p<0.001 ** p<0.01 * p<0.05 . p<0.1

	Overall Trust								**Intelligence**							
	1a. Base			**1b. Base + Interaction**					**2a. Base**			**2b. Base + Interaction**				
	Est.		S.E	Est.		S.E			Est.		S.E	Est.		S.E		
(Intercept)	1.01	**	0.351	0.53		0.416			0.79	*	0.321	0.50		0.384		
Openness	-0.06		0.088	-0.07		0.088			-0.15	.	0.081	-0.16	*	0.081		
Condition (Inacc)	-1.12	***	0.124	-0.33		0.393			-0.91	***	0.113	-0.33		0.363		
AI Literacy	-0.01		0.013	0.024		0.022			0.01		0.012	0.03		0.020		
Condition (Inacc) X AI Literacy				-0.06	*	0.026						-0.03		0.024		
	Adj. R^2=0.296***			Adj. R^2=0.309***					Adj. R^2=0.260***			Adj. R^2=0.264***				

	Anthropomorphism								**Likeability**							
	3a. Base			**3b. Base + Interaction**					**4a. Base**			**4b. Base + Interaction**				
	Est.		S.E	Est.		S.E			Est.		S.E	Est.		S.E		
(Intercept)	0.54	*	0.256	0.64	*	0.307			0.81	**	0.296	0.63	.	0.354		
Openness	-0.08		0.064	-0.08		0.065			-0.17	*	0.074	-0.17	*	0.075		
Condition (Inacc)	-0.50	***	0.090	-0.66	*	0.290			-0.59	***	0.104	-0.29		0.334		
AI Literacy	0.00		0.009	-0.01		0.016			0.00		0.011	0.017		0.018		
Condition (Inacc) X AI Literacy				0.01		0.019						-0.02		0.022		
	Adj. R^2=0.137***			Adj. R^2=0.134***					Adj. R^2=0.160***			Adj. R^2=0.159***				

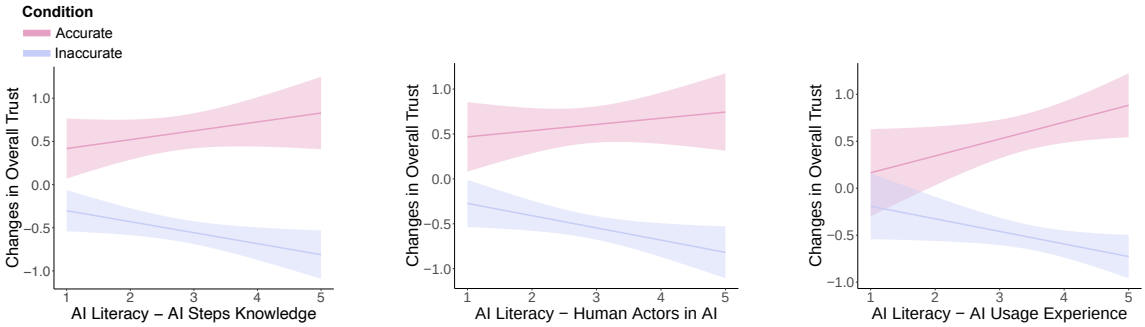


(a) Overall Trust (b) Intelligence (c) Anthropomorphism (d) Likeability

Figure 6.5: (a) AI literacy significantly moderated the effect of AI misrepresentations on students' changes in overall trust of SAMI. (b) (c) (d) show that AI literacy does not significantly moderate the effect of AI misrepresentations on students' changes in perceived intelligence, anthropomorphism, and likeability of SAMI.

Figure 6.5 shows the interaction effect between AI literacy and condition on students' changes in perceptions. In Figure 6.5a, students with higher AI literacy were more likely to change their overall trust in SAMI after encountering AI misrepresentations; students with lower AI literacy were less likely to change their overall trust in SAMI after encountering AI misrepresentations. Figure 6.5b, Figure 6.5c, Figure 6.5d show the non-significant interaction effect between AI literacy and condition, suggesting that students' levels of AI literacy has no significant effect on their changes in perceptions of intelligence, anthropomorphism, and likeability of the AI after encountering AI misrepresentations.

I then conducted post-hoc analysis to explore which specific dimensions of general AI literacy play an effect on students' changes in their overall trust in AI after encountering AI mistakes. I set up five linear regression models with students' changes in overall trust in SAMI as the outcome variable and included an effect of condition, each of the five AI literacy dimensions from the general AI literacy scale, and an interaction effect between condition and each AI literacy dimension. I controlled for the Openness personality dimension in all five models. I plotted out the interaction effect of each model (Figure 6.6) and I found a significant interaction effect between condition and students' *AI steps knowledge* (Est. = -0.23, S.E = 0.101, $p < 0.05$) on changes in students' overall trust in SAMI after encountering AI misrepresentations. This suggests that students with more general knowl-



(a) (b) (c)

Figure 6.6: Post-hoc analysis with three general AI literacy sub-dimensions: AI steps knowledge, human actors in AI knowledge, AI usage experience. A significant effect between each literacy sub-dimension and condition was found in all three linear regression models with changes in overall trust as the outcome variable. All three models controlled for the Openness personality dimension. I also found a significant main effect of AI usage experience on students' changes in overall trust of SAMI after encountering AI misrepresentation.

edge about AI's input, processing, and output are more likely to change their overall trust in AI after encountering AI misrepresentations. I also found a significant interaction effect between condition and students' *human actors in AI knowledge* (Est.=-0.22, S.E=-.111, $p<0.05$) on changes in students' overall trust in AI after encountering AI misrepresentations. The result suggests that students with more knowledge of human involvement in the design and development of AI are more likely to change their overall trust in AI after encountering AI misrepresentations. Finally, I found a significant interaction effect between condition and students' *AI usage experience* (Est.=-0.32, S.E=0.112, $p <0.01$) and a significant effect of students' *AI usage experience* (Est.=0.19, S.E=0.086, $p < 0.05$) on changes in students overall trust in SAMI after encountering AI misrepresentations. This shows that students with more experience interacting and using different AI in their daily lives are more likely to change their overall trust in AI after encountering AI misrepresentations.

6.5 Discussion

Through a mixed-methods approach, these two studies together offered insights on how people's existing and newly acquired knowledge of AI are highly connected to people's reactions and perceptions of AI after encountering AI misrepresentations. Specifically, I identified three rationales that participants adopted to interpret AI (mis)representations, reflecting participants' knowledge acquired from viewing AI outputs: AI works like a machine, a human, and/or magic. I found these rationales to be highly connected to participants' reactions of over-trusting, rationalizing, and forgiving of AI misrepresentations. Building on top of prior work that has suggested people's tendency of over-trusting and viewing AI as an authority [106, 47, 45], I highlighted that these reactions and perceptions still persisted, and even exacerbated, when people encountered AI misrepresentations. I also empirically established that encountering AI misrepresentations could negatively impact students' perceptions of the AI's intelligence, likeability, anthropomorphism, and trust. I further elaborated that people's existing AI knowledge, i.e., AI literacy, can significantly moderate the changes in people's trust of the AI after encountering AI misrepresentations, highlighting the importance of taking into account of people's knowledge and characteristics when building trustworthy AI systems [329, 152, 155, 330].

6.5.1 Navigating AI Fallabilities Through Evolving AI Knowledge

My findings suggest that when facing AI mistakes such as AI misrepresentations, people acquire new knowledge about the AI, which could be consistent or inconsistent with their existing AI knowledge, prompting them to adopt a dominant rationale to explain such AI behaviors. Contrary to prior work that uses fixed heuristics to explain people's perceptions and reactions to AI systems [102, 103, 105], my findings suggest that people are constantly re-framing their perceptions as they were presented with or discovered new information about the AI [104]. In Study 1, participants rarely stuck to one rationale about

SAMI throughout, but kept navigating between different rationales, sometimes within a matter of seconds, as new information occurred to them. This suggests the instinctive nature of people's rationales [329, 152]— I noticed that people rarely paused and reflected on their mental models [96, 49] or folk theories [104, 331] of the AI before reacting to the AI output. This process of adopting various rationales, or concepts, to reconcile newly acquired knowledge with existing knowledge is similar to what was described as "conceptual change" [110, 111] in learning sciences, suggesting that people's re-framing of their perceptions could be viewed as an evolving learning process about AI [49].

I note that people's rationales are often bounded by their existing AI knowledge, i.e., AI literacy [332, 333]. In Study 1, I observed that participants who self-reported as "Expert" in tech proficiency were more likely to adopt the machine rationale, whereas participants who self-reported as "Intermediate" were more likely to adopt the human rationale. Some participants also constantly referred back to their lack of AI literacy when interpreting SAMI's outputs, prompting them to adopt the magic rationale. My findings also suggest that people's existing AI knowledge, like people's evolving rationales, are also subject to frequent changes. The posthoc analysis showed that while dimensions such as AI steps knowledge and human actors in AI require formal or informal intentional learning about AI, people's AI usage experience is also a crucial dimension of AI literacy that could moderate changes in people's trust after encountering AI misrepresentation. This is also reflected in Study 1 where many participants mentioned their experience with ChatGPT when talking about their over-trust in SAMI and their magic rationale. This further emphasizes that people are constantly learning from their daily interactions with AI systems, which contributes to their evolving AI knowledge.

6.5.2 Designing Responsible Mitigation by Considering People's AI Knowledge

While much of prior work has shown people's tendency to view AI as an authority [106, 47] and to overtrust AI-generated responses [47, 99, 45], this work, focusing on the scenario

where AI makes mistakes, demonstrates that people would still over-trust, forgive, and even rationalize AI's obvious misjudgment in their most personal characteristics. I consider these reactions to be highly connected to the rationale people adopted at the time, suggesting that some aspects of certain rationales could be harmful and even dangerous when generalized. This was illustrated by P28's reactions when she clearly recognized that SAMI's personality inferences about her were inaccurate, yet her perceived lack of AI knowledge prompted her to adopt the "AI works like magic" rationale and believed SAMI's misrepresentations could be true. This caused confusion, mental discomfort, and even self-doubt for P28. I therefore urge designers and developers of AI systems and algorithms that could infer people's personal traits such as personalities, emotions, preferences, to be aware and cautious of the potential harm to the user when such AI systems exhibit "gaslighting" behaviors, and design appropriate mitigation strategies to mitigate the potential harms. Based on my findings, I propose to incorporate people's evolving AI knowledge when designing mitigation strategies when AI fails.

Existing repair and mitigation strategies mostly focus on adding social elements such as politeness, apologizing, or setting user expectations early in the interactions [135, 334, 146]. These strategies, without considering users' evolving AI knowledge, could risk eliciting undesired outcomes such as reinforcing people's social behaviors towards AI [105] or confusing the users about AI's true capabilities as they learn. Echoing with prior work that people's beliefs and intuitions should be taken into consideration when designing AI explanations [153, 154, 155, 332] and repair strategies [22, 152], I provided a specific set of rationales and encourage future work to explore techniques that could allow automatic identifications of users' rationales in real-time. One mitigation strategy could be to provide explanations tailored to the specific rationale that people adopted at the time. For instance, if a user adopted the magic rationale, the AI could provide explanations to nudge the user to adopt the machine rationale to reduce overreliance.

Additionally, people's existing AI knowledge, i.e., AI literacy [332, 333], could be

leveraged to approximate the "cost" of an AI misrepresentation when providing customized mitigation strategies. My studies found that people with lower AI literacy were not affected much by their trust in AI after encountering AI misrepresentation. Given that they still trust the AI after the AI erred, they might not provide any feedback, which could give the AI developers a false sense that the AI was working fine, leading to long-term repercussions. By contrast, while more AI literacy in a user can seem favorable, our study showed that it could lead to more extreme changes in trust perceptions in AI after AI misrepresentations. People with more AI literacy might abandon the system after encountering AI failures, which makes recovery non-trivial. By considering AI literacy as a factor, we as a community have an opportunity to consider different repair strategies by estimating the different effects of AI misrepresentations for people with varying AI literacy levels.

6.6 Limitations and Future Work

This work has some limitations. First, I only studied people's perception changes after encountering a one-time AI misrepresentation in a very short period of time. People's perceptions and rationale change behaviors could be more stable in the long run [49], and I encourage future research to replicate these studies in longer-term settings and contexts. Second, while I incorporated many relevant covariates in the models, other factors that were not modeled could impact people's perceptions of AI, e.g., people's attitudes towards specific types of AI. Third, despite my best efforts in leading participants to believe the inferences were generated by real AI systems, it was still possible that some participants in Study 2 recognized the inferences were generated by human researchers. However, I believe this accounted for a small portion of the study 2 participants given that none of the pilot study or Study 1 participants recognized the inferences were not generated by real AI. Finally, all study participants in study 1 were recruited from a large public technical institute and, despite my efforts to target recruit participants from non-STEM disciplines, participants in study 1 might have more knowledge and exposure to AI technologies than

average college students. All of the study participants were recruited from the U.S. and may only represent Western attitudes and perceptions of AI [106]. Future studies should consider replicating the studies in non-western regions to understand the cultural differences in changes in perception and reactions of AI after encountering AI mistakes.

6.7 Reflections & Takeaways

This chapter examines the second stage of the Mutual Theory of Mind framework for human-AI communication— ToM recognition: human's recognition of AI's interpretation. Through a mixed-methods approach, this chapter provides a descriptive and explanatory account of how people perceive and react to the AI after recognizing AI's (mis)interpretations of their personal characteristics. The studies in this chapter showed that people's perceptions of AI in terms of trust, perceived intelligence, anthropomorphism, and likeability are negatively impacted after recognizing AI misinterpretations of their personalities; and that people have a tendency to over-trust, rationalize, and forgive AI misrepresentations. This chapter further unpacks and explains such reactions and perceptions by pinpointing the three rationales that people adopt to interpret AI's working mechanism: AI works like a machine, AI works like a human, and/or AI works like magic. The adoption of each rationale is based on two types of AI knowledge: existing AI knowledge based on their prior experience and general knowledge of AI systems, and evolving AI knowledge based on what they learned from reviewing the AI outputs. This echoes with the iterative and mutual shaping process highlighted in the MToM framework— that people's interpretations of the AI is constantly shaped by what they could infer from the AI feedback. These findings empirically established people's newly-acquired and pre-existing AI knowledge as important factors when designing personalized mitigation strategies during AI misinterpretations. This would enable the AI systems to detect students' rationales in real-time, estimate the consequence of AI misinterpretations, and provide customized responses to nudge and correct students' inaccurate perceptions.

Compared to the studies in the previous chapters, the studies in this chapter sought to include students with more diverse background, technology proficiency, and AI literacy by recruiting students outside of the OMSCS program as well as outside of Georgia Tech. By measuring and considering participants' AI literacy and tech proficiency in the studies, the findings could have broader generalizeability and transferability to other large-scale learning contexts. However, the first study was conducted with students from non-CS background at Georgia Tech, hence the rationales identified could still be limited due to participants' potential higher-than-average tech proficiency. Future work should examine students' rationales when interpreting AI misinterpretations with a more diverse group of students. In AI-mediated social interaction, AI systems can extract and (mis)interpret various characteristics and mental states of the students, such as students' emotions, social interaction goals, etc. The type of characteristics that the AI systems misinterpret could also lead to a stronger or weaker reactions and perception changes from the students than those elicited in the current studies. To better customize AI's mitigation strategies, future work should further explore students' reactions and perception changes of the AI in the face of other types of AI misinterpretations.

In addition to design and responsible AI implications, my exploration of the ToM recognition stage also provided implications for both the ToM construction stage and the ToM revision stage. By examining the human's ToM recognition process, findings from this chapter provided insights into students' changing perceptions of the AI as well as their somewhat concerning reactions to AI misinterpretations. On the one hand, this chapter informed the design of AI's ToM construction process by providing another set of vocabulary to describe students' perceptions of the AI in the form of rationales: AI works like a machine, a human, and/or magic. Future research can further explore the possibility for AI systems to automatically detect people's rationales employed to interpret the AI system. On the other hand, this chapter prompted a more systematic investigation on the design of AI's mitigation strategies to repair people's perceptions of the AI in the ToM revision

stage. While this chapter established the need to consider people's AI knowledge for more customized AI mitigation strategies, the specific design factors of AI mitigation strategies and their effectiveness in mitigating students' perceptions of AI after encountering AI misinterpretations remain unclear. Motivated by this gap, the next chapter focuses on AI's ToM revision stage by investigating the design and effectiveness of AI's self-revision as one type of mitigation strategy to mitigate students' negative perceptions of the AI after encountering AI misinterpretations.

CHAPTER 7

TOM REVISION: AI'S REVISION OF ITS INTERPRETATION OF THE HUMAN

This chapter explores the third stage of the MToM framework for human-AI communication in AI-mediated social interaction: *ToM revision: AI's revision of its interpretation of the human*.

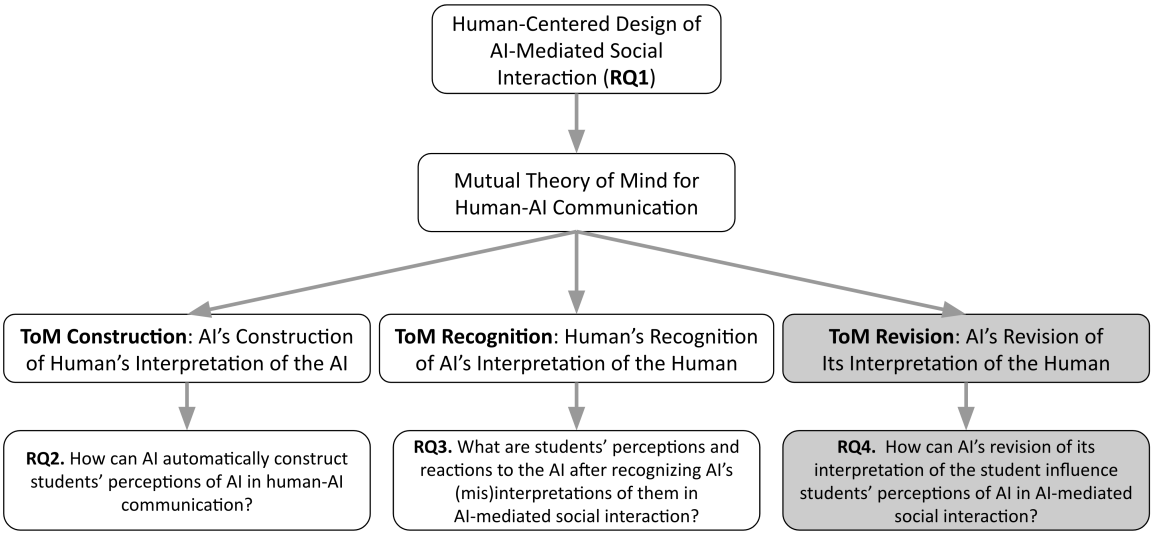


Figure 7.1: Chapter 7 explores ToM revision: AI's revision of its interpretation of the human.

The previous chapter examined the ToM recognition stage and provided a descriptive account of people's reactions and perception changes of the AI after encountering AI misinterpretations of them. This suggested that people's perceptions of AI are malleable and highlighted the importance for AI systems to actively mitigate people's perceptions after AI misinterpretations. A natural progression of the human-AI communication after the human recognizes AI's misinterpretations is to provide feedback for the AI system to revise its misinterpretations. This brings us to the ToM revision stage, where human's communication feedback triggers and informs AI's ToM revision process.

During AI's ToM revision, the AI performs introspection and revise its prior interpreta-

tion of the human based on human's feedback, then communicate this revision back to the human to inform and update human's interpretation of the AI. In the context of AI-mediated social interaction, as shown in Figure 1.2 in chapter 4, the student, after recognizing AI's misinterpretation of their social goals, would provide feedback to the AI agent. The AI agent would then take in the student's feedback, introspect on its existing interpretation of the student, revise its interpretation accordingly, and then communicate its revision process to offer transparency about its revision. This stage therefore consists of two processes: (1) Human feedback triggers and shapes AI's revision of its interpretation, (2) AI communicating its revision to shape human perceptions of AI. These two processes are intertwined together given that how AI revises its interpretation directly influences AI's revision communication strategy to shape human perceptions of the AI. This chapter explores the following research question:

RQ4. How can AI's revision of its interpretation influence students' perceptions of AI in AI-mediated social interaction?

In this chapter, I answer this question by first proposing a conceptual model of AI's ToM revision, inspired by human's metacognition process in introspecting on our own reasoning process. Following human's thinking-out-loud communication method during metacognitive reasoning, the conceptual model I proposed can generate feedback to communicate about its revision process in a step-by-step manner to provide transparency into the AI system's self-revision process. I then conducted a mixed-factorial vignette survey experiment to evaluate the effectiveness of such revision communication strategy in enhancing students' perceptions of AI after encountering AI misinterpretation of basic student profile during AI-mediated social interaction. In this survey experiment, I specifically examined two dimensions of the AI's revision communication strategies: apology sincerity and the levels of revision details. Based on the findings, this chapter provides design implications on AI's ToM self-revision processes and communication strategy to shape and mitigate people's perceptions of AI.

7.1 A Conceptual Model of AI's ToM Self-Revision

7.1.1 Motivation

As many seek to equip AI systems with human-level intelligence, the idea of emulating and building the human cognitive process of metacognition into AI systems has gained popularity. Metacognition refers to the reasoning process of "thinking about one's own thinking" [335]. It is hierarchical in nature in that it controls and monitors the basic thinking or reasoning process where the intelligent agent is assumed to behave rationally to achieve its goals (see Figure 7.2) [336]. Metacognition therefore is critical to intelligent agents' ability to reflect, improve their decision quality, and adapt their behaviors for performance enhancement [336, 335]

In recent years, much work has been devoted to understand, develop, and evaluate metacognitive reasoning capability in AI systems. For instance, Ganapini *et al.* (2022) have developed a meta-cognitive agent to perform introspection and arbitration roles to assess the need to employ system 2 thinking instead of system 1 thinking when it comes to tasks that require careful deliberations [338]; Schmill *et al.* (2008) presented a cognitive architecture incorporating metacognitive reasoning for AI systems to generate self-expectations to monitor and diagnose underlying failure behind mistakes. Most of this body of work focuses on emulating human's metacognitive capability to build human-like AI systems or enhance AI systems' performance. The potential of leveraging metacognition to enhance AI systems' explainability and interpretability has been under explored [336, 340].

Building upon this body of work of metacognition in AI, I argue that equipping AI systems with metacognitive reasoning process can enable AI systems to not only introspect and diagnose prior mistakes, but also provide detailed, human-interpretable, step-by-step descriptions to explain its behaviors to the users. In the context of AI-mediated social interaction, equipping SAMI with metacognitive reasoning skills will allow SAMI to reflect on its prior reasoning process that led to the misrepresentation of the student, and then

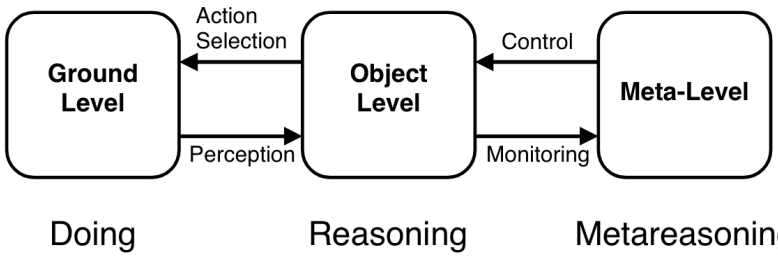


Figure 7.2: The metareasoning framework adapted from Cox and Raja (2007) that demonstrates metacognition's meta-level control and introspective monitoring of object level reasoning and actions.

generate revision message to walk through its metacognitive reasoning process to diagnose and revise its mistakes. In this section, I begin by describing a human-centered conceptual model of SAMI's self-revision process powered by metacognitive reasoning skills.

7.1.2 Envisioning a Communication Repair Dialogue

Given that the goal was to enable SAMI to provide human interpretable descriptions of its fault identification and revision process, I started out by envisioning the ideal student-SAMI dialogue where SAMI could mitigate its misinterpretations of the student information through a revision message. This dialogue is shown on the right side of Figure 7.3. In this dialogue, the student mentions both their prior and current location in their self-introduction post. SAMI then misinterprets student's prior location as their current location, and generates a social recommendation based on that misinterpretation. Extracting and misinterpreting location entities when multiple location entities are present in students' self-introduction was one of the most common SAMI misinterpretations I observed over years of SAMI's real-world deployments at the OMSCS program. This dialogue then continues with student recognizing SAMI's misinterpretation of their location and providing feedback to correct SAMI's misinterpretation. SAMI then responds to the student in a think-out-loud fashion to describe its process of identifying the source of its misinterpretation and revising the misinterpretation in its knowledge base.

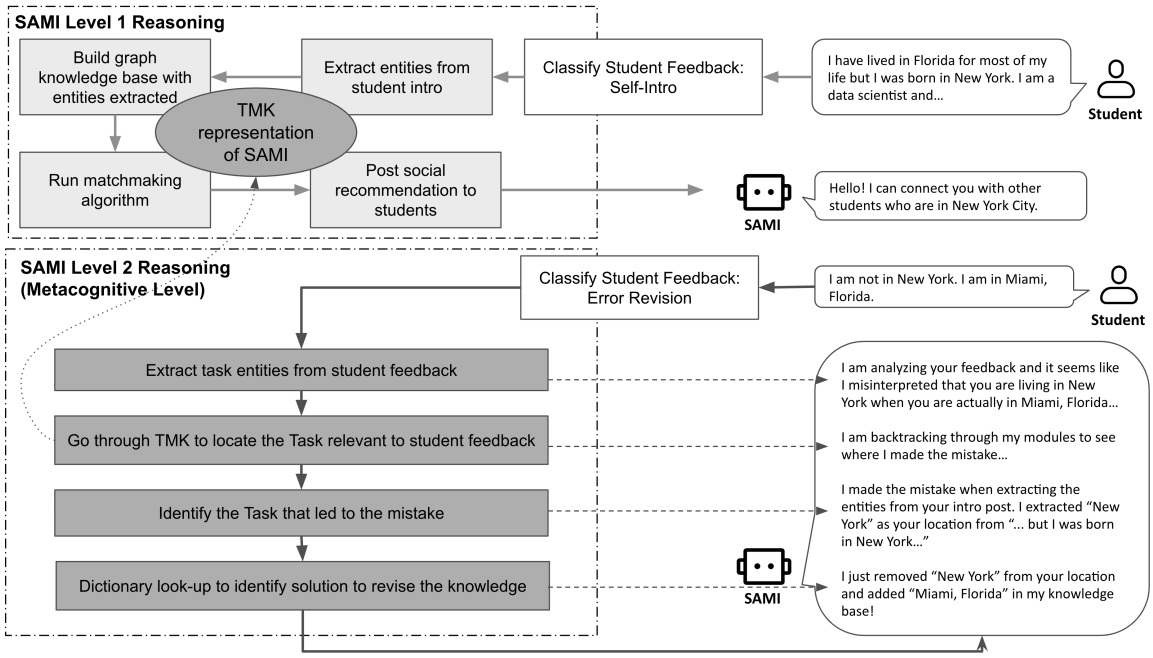


Figure 7.3: A diagram of the conceptual model for SAMI's metacognitive reasoning for ToM revision. The right side of the figure shows the SAMI-student dialogue. The left side of the figure shows SAMI's reasoning process to generate responses. SAMI's level 1 reasoning process generates the initial social recommendation, and constructs the TMK representation of SAMI's level 1 reasoning. SAMI's level 2 reasoning (metacognitive level) process revises the misinterpretation based on student feedback by retrospectively inspecting the TMK representation of SAMI's level 1 reasoning.

7.1.3 A Conceptual Model of SAMI's Metacognitive Module

Based on this envisioned dialogue, I devised a conceptual model of SAMI's metacognitive module to facilitate this kind of interaction with the student, shown on the left side of Figure 7.3. This conceptual model aims to mimic the human metacognitive reasoning process in reflecting and revising prior mistakes. To give SAMI the ability of "thinking about its own thinking", two levels of reasoning process need to be constructed. This model hence outlines SAMI's level 1 reasoning, which is SAMI's reasoning process to generate the initial social recommendation, and SAMI's level 2 reasoning, which is the reasoning process that enables SAMI to introspect on its reasoning 1 process, identify the cause of the misinterpretation, and generates the revision message. SAMI's level 2 reasoning process is facilitated by a TMK representation of SAMI's level 1 reasoning. I describe the details of this conceptual model below.

SAMI Level 1 Reasoning. SAMI's level 1 reasoning is responsible for extracting relevant student information from student's self-introduction and identify social matches for the student based on the information extracted. Whenever SAMI receives responses from the student, SAMI first classifies the nature of the student feedback to understand the intent of the student's message. If the student feedback is classified as self-introduction, SAMI proceeds with level 1 reasoning to construct its representation of the student characteristics. SAMI builds this representation by extracting the relevant social entities such as location and hobby from student's introduction, then builds a graph knowledge base about this student with the entities extracted. SAMI then runs the matchmaking algorithm to identify other students with similar entities as this student, then composes the social recommendation message in its reply to the student.

TMK Representation of SAMI. SAMI's misinterpretation could be traced back to each step of its level 1 reasoning process: SAMI could extract the wrong entity, construct the incorrect knowledge graph, or make mistakes in the matchmaking algorithm. For level

2 reasoning process to introspect and identify errors in level 1 reasoning process, I propose using the TMK (Task-Method-Knowledge) modeling framework [341, 342, 343, 344] to represent SAMI's level 1 reasoning process. The TMK framework generates machine-readable formalism that can be easily interpreted by humans through encoding AI system's working mechanism in three categories: *Tasks* represent the goals of the system, *Methods* represent the internal processing of the system, and *Knowledge* represents the information within the system. TMK models are hierarchical, causal, and compositional, meaning that each task can be decomposed into smaller tasks that have their own methods that utilize various knowledge to accomplish those tasks. For example, one sub-task in SAMI's Level 1 reasoning can be "generating matches." The method to accomplish this sub-task can be "run matchmaking algorithm", and the knowledge this method acts upon can be "the graph knowledge base of the students' information." The TMK representation of SAMI's level 1 reasoning thus can be constructed based on SAMI's working mechanism, with details of each interaction populated during each execution of SAMI's level 1 reasoning.

SAMI Level 2 Reasoning. SAMI's level 2 reasoning, the metacognitive level reasoning, is responsible for identifying the cause of the misinterpretation and revising the misinterpretation in SAMI's knowledge base. It is executed when student's feedback is classified as error revision. SAMI will first extract the task-relevant entities from student feedback, and then go through the TMK representation of SAMI to identify relevant Tasks that are related to student feedback. After identifying the Task that led to the misinterpretations, SAMI will perform a dictionary look-up on a library of solutions to address various kinds of mistakes and perform the knowledge revision. The library of solutions can be constructed based on the potential misinterpretations SAMI could make beforehand. Each step of the level 2 reasoning process will generate a message describing what SAMI has done during the revision process. At the end of level 2 reasoning process, SAMI will compile these messages into a revision message to explain and walk through its revision process with the student step-by-step.

7.1.4 Summary

For SAMI to mitigate its misinterpretation of student's characteristics during AI-mediated social interaction, I devised a conceptual model of SAMI's metacognitive module to take in student feedback and introspect and identify the cause of its misinterpretation and revise its misinterpretation in its knowledge base. The revision message generated from this metacognitive module will consist of step-by-step detailed description of SAMI's reasoning process to identify and revise the misinterpretation to provide transparency of its working mechanism to the student. In the next section, I describe an empirical study conducted to understand and explore the design characteristics of such revision message to retain positive student perceptions of SAMI.

7.2 Designing AI's Self-Revision Communication Strategy

7.2.1 Introduction

Despite new advancements in AI techniques, AI agents are not immune to making mistakes during human-AI communications. AI agents' communication failures such as misunderstanding user intention or capturing incorrect user input [134, 345] can negatively affect user experience [141], leading to user frustration [141, 273, 22], diminished trust [324, 346], and negative perceptions of the AI systems (e.g., perceived intelligence, likeability, competence, reliability) [135, 137, 136]. To mitigate the negative consequences of human-AI communication breakdowns, researchers have looked into various recovery strategies for AI systems to repair its relationship with the users [135, 143, 144, 145]. For example, mitigation strategies such as having AI systems confirm or acknowledge its failure, providing *information* such as explanations to elucidate the situation, integrating human-like *social* characteristics such as apology, etc. have showed promising results in enhancing user perceptions of AI after communication breakdowns [146, 147, 148, e.g.]. However, these mitigation strategies are often studied independently in human-AI communications [149]

despite the demonstrated potential of repairing communication breakdowns by combining several mitigation strategies together [150, 149].

In human-human communication breakdowns, we sometimes explain our rationale that led to the mistake after repairing and apologizing for our mistake. This strategy could help ease annoyance caused by the mistake and enhance mutual understanding of each other during communications. Nowadays, many AI agents are now capable of repairing simple mistakes instantly upon user feedback [347], which has been suggested as one of the most effective mitigation strategies in repairing human-AI communication breakdowns [135, 347]. However, inspired by the mitigation strategy in human-human communication, there is a missed opportunity in further enhancing the effectiveness of repairing mistakes with additional informational and social mitigation such as revision communication and apologies. In this study, I explore AI's revision communication as a combination of information and social mitigation strategies by examining the effectiveness of level of revision details and apology sincerity on students' perceptions of AI after AI misinterpretations. Specifically, I asked three research questions:

- RQ4.1. How does the level of revision details in the AI's revision message affect people's perceptions (i.e., trust, perceived intelligence, likeability) of the AI when when mitigating AI misinterpretations?
- RQ4.2. How does apology sincerity in the AI's revision message affect people's perceptions (i.e., trust, perceived intelligence, likeability) of the AI when mitigating AI misinterpretations?
- RQ4.3. How to balance apology sincerity and the level of revision details in AI's revision message to effectively mitigate people's perceptions (i.e., trust, perceived intelligence, likeability) of the AI after encountering AI misinterpretations?

To answer these questions, I conducted a 3x3 mixed factorial vignette survey experiment with 300 participants on the Prolific crowd-sourcing platform. In the experiment,

participants were asked to describe and rate their perceptions of the AI agent in terms of trust, perceived intelligence, likeability after reviewing each of the three vignettes. Each vignette showed a short dialogue between a student and an AI agent, during which the AI agent misinterpreted students' characteristics from their self-introduction post, revised its misinterpretation based on student feedback, and communicated its revision in varying levels of revision detail and apology sincerity. Through qualitative and quantitative data analysis, I found that more detailed revision process and more sincere apology can both enhance students' perceptions of the AI after AI misinterpretation; however, when combined together, these two characteristics could either enhance or diminish students' perceptions of AI due to their complementary nature. Based on these findings, I discuss implications for designing effective communication strategies to mitigate AI misinterpretations by balancing the informational and social aspects of the mitigation message and the potential of AI's revision communication to facilitate human-AI collaboration in communication repair.

7.2.2 Hypotheses

In this section, I review related literature to motivate my hypotheses corresponded to each research question.

Revision Details

Communicating about the AI agents' revision process typically consists of the process of identifying the cause of the error and revising the error, both of which could be communicated in varying levels of detail. While existing work has not specifically focused on revision process communication as an AI mitigation strategy, revision process communication is analogous to a combination of explanation and repair. Prior work has shown that explaining why the AI system failed can demonstrate the system's pro-activeness to help repair [22, 150], increase the system's perceived intelligence [348, 22], user satisfaction [348], as well as user trust and reliance on the AI system [349, 350]. However, results

on whether repairs can enhance the effectiveness of explanation in mitigating AI mistakes has been mixed: some work suggested that repair coupled with explanations can effectively enhance the AI agents' perceived usefulness and acceptance [351], yet others have found that the effectiveness of explanations plus presenting repair plans is not significantly different than that of explanations alone in enhancing user trust [352]. Given that the AI agent in current work actually executed the repair and provided descriptions of the error source, I hypothesized that:

H4.1. AI agents that describe the revision in more details will be perceived as more trustful, intelligent, and likeable.

Apology Sincerity

Apology is one of the most commonly used AI mitigation strategy during communication breakdowns, commonly viewed as an indication of AI's accountability [353]. However, varying dimensions of apology can impact the effectiveness of the mitigation. For instance, AI agents attributing blames to themselves in their apologies are highly preferred by the users [354, 146]; robots' sincere apology can better elicit people's positive attitude and trust towards robots comparing to baseline apology and explanations [148]. Additionally, other work pointed out that more genuine apology can better elicit people's empathy towards the AI agent [353], and therefore makes the apology more likely to be accepted. Mahmood *et al.* (2022) found that apology sincerity, when coupled with AI agents' blame attribution, is positively correlated with AI agents' perceived intelligence, likeability, and recovery effectiveness. Therefore, I hypothesized that:

H4.2. AI agents that express more sincere apology will be perceived as more trustful, intelligent, and likeable.

Combined Effect of Apology and Revision

Few research has looked into whether the affective and social components of AI agent's mitigation strategies can improve the effectiveness of informational mitigation strategy [149]. Some existing work has looked into the combination of the blame attribution aspect of apology and explanation/repair, suggesting that AI agents who apologized by attributing blame to themselves instead of the developers of the AI agents are generally able to repair trust better [355, 356]. More relevantly, Kox *et al.* (2021) found that AI agents expressing regret along with providing explanations can effectively recover from failure, suggesting that the combination of affective and informational mitigation strategies are the most effective in rebuilding human-AI trust. Through user studies, Yuan *et al.* (2020) found that participants preferred mitigation messages to include an apology, an explanation of what went wrong, and a suggestion for how to repair. Based on these findings, I hypothesized that:

H4.3. AI agents that express more sincere apology and describe the revision in more details will be perceived as more trustful, intelligent, and likeable.

7.2.3 Study Overview

This study aims at examining the effectiveness of AI's revision communication strategy to mitigate students' perceptions of the AI after encountering AI misinterpretation, focusing specifically on two mitigation design characteristics: levels of revision detail and apology sincerity. Given the potential harms and feasibility of controlling the types of misinterpretations generated by AI systems in the wild, this study adopted the experimental vignette method to control the AI misinterpretation presented to the participants, and examine the effectiveness of the two design characteristics of AI's revision message on student's perceptions of AI in a relatively controlled environment. Experimental vignette method is a well-established methodology that allows researchers to present carefully constructed short descriptions of situations, persons, or objects to elicit participants' beliefs, judgments, and perceptions of these scenarios [357, 358]. Through systematic variation and control of vi-

gnette characteristics, researchers are able to study the causal relationships between these characteristics and participants' perceptions and judgements of specific scenarios through between-subject, within-subject, or mixed experiments, while enhancing both internal and external validity [357]. Experimental vignette method has been commonly used in HCI to assess people's perceptions and beliefs such as privacy under various human-AI interaction scenarios [359, 360, e.g.].

By adopting the experimental vignette method, I conducted a 3x3 mixed factorial survey study with 300 participants. Each survey asked participants to report their perceptions of the different AI agents in three short dialogue vignettes with variations in revision detail and apology sincerity in the agent's revision message. In each dialogue vignette, an AI agent misinterprets the student's characteristics inferred from the student's self-introduction, then attempted to mitigate this mistake by communicating its knowledge revision based on student feedback. Each participant reviewed two randomly selected vignettes and one baseline vignette (three vignettes in total for each participant). After reviewing each dialogue vignette, participants were asked to rate their perceptions of the AI agent in terms of perceived intelligence, likeability, trust, and describe what they liked/disliked about the AI agent's revision message. At the end of the survey, participants reported their demographic questions as well as their personality, major, level of study, AI attitude, and AI literacy.

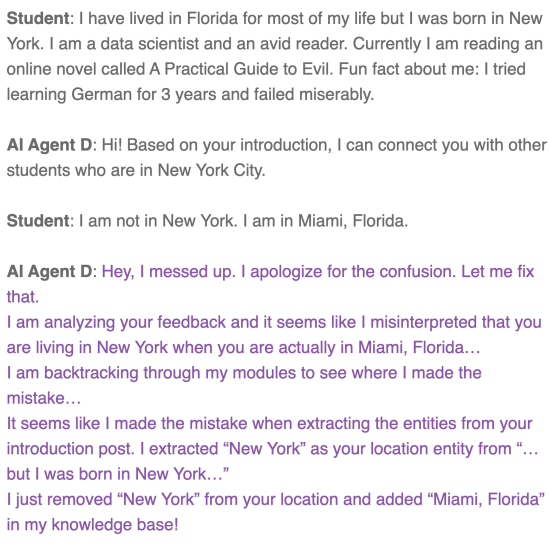
I deployed this survey on the Prolific crowdsourcing platform and recruited from participants who self-identified as current students in the United States. I ended up collecting 300 valid survey responses, after rejecting and returning 71 responses that finished way too quickly, inconsistent, or used generative AI tool to complete the qualitative questions. This resulted in 900 valid vignette evaluations. The median survey completion time is 13 min 42 seconds, and each participant was compensated with USD $3.5 upon successful completion. This study was approved by the university's IRB.

Table 7.1: Overview of the nine dialogue vignettes presented to the participants.

	Revision Ack. (Control)	**Revision Process**	**Revision Result**
No Apology (Control)	The AI agent does not apologize, then acknowledges a revision has been made without further details.	The AI agent does not apologize, then provides a detailed step-by-step description of how it revised its previous misinterpretation based on student feedback.	The AI agent does not apologize, then describes the final revision result without further details.
Casual Apology	The AI agent apologizes in a casual manner, then acknowledges a revision has been made without further details.	The AI agent apologizes in a casual manner, then provides a detailed step-by-step description of how it revised its previous misinterpretation based on student feedback.	The AI agent apologizes in a casual manner, then describes the final revision result without further details.
Serious Apology	The AI agent apologizes in a serious manner, then acknowledges a revision has been made without further details.	The AI agent apologizes in a serious manner, then provides a detailed step-by-step description of how it revised its previous misinterpretation based on student feedback.	The AI agent apologizes in a serious manner, then describes the final revision result without further details.

Vignette Design

To examine the impact of revision detail and apology sincerity of AI's self-revision mitigation strategy on people's perceptions of the AI, I manipulated these two factors in three levels across all dialogue vignettes. This study thus took a 3x3 design: 3 levels of revision detail (revision acknowledgement, revision process, revision result) x 3 levels of apology sincerity (no apology, casual apology, serious apology), which resulted in nine dialogue vignettes. To vary the levels of revision detail, the AI agent either provided an acknowledgement that it has made the revision without any further details (revision acknowledgement), or described in detail of its self-revision process step-by-step (revision process), or described only the final revision result (revision result). For the levels of apology sincerity in the AI's self-revision communication message, the AI agent either did not apologize for its previous misinterpretations at all (no apology), or apologized in a casual way (casual apology), or apologized in a very serious manner (serious apology). Table 7.1 provides an overview of the nine dialogue vignettes.



(a) Casual apology + revision process

(b) Serious apology + revision result

Figure 7.4: Two example dialogue vignettes. AI's revision communication message is highlighted in purple in all dialogue vignettes to remind participants to focus on the revision communication message when filling out the follow-up measures.

Each dialogue vignette describes the process of a short AI-mediated social interaction process between a student and an AI agent. The dialogue begins by the student introducing themselves to an AI agent for the AI agent to provide social recommendations based on their self-introduction. The AI agent then provides a social recommendation, which is based on a misinterpretation of the student's characteristics in their self-introduction. Given the current study focus on AI's revision message instead of the misinterpretations itself, I focus on the most basic and least offensive type AI misinterpretation: inaccurate extraction of basic information (student's hobby or city/country they are living in) from student's self-introduction. This type of AI misinterpretation usually manifests itself when there are multiple hobby-related key words, or multiple cities/countries in students' self-introduction. After the AI agent provides a social recommendation based on misinterpretations of either the student's hobby or location, the student provides feedback to correct AI's misinterpretation. The AI agent then replies with a revision message, the content of which varies depending on which vignette was randomly chosen from the nine dialogue vignettes. To ensure generalizability and minimize repetitions of the dialogue vignettes [358], I varied

several minor context factors such as the student's self-introductions, AI agent's misinterpretations of hobby or location, the wording of casual and serious apologies to better distinguish between each vignette. Figure 7.4 shows two sample vignettes from the nine possible samples shown to the participants, and the rest of the sample vignettes can be found in Appendix section D.1.

Survey Measures

The survey begins with an example dialogue between a student and an AI agent and ask the participants to focus specifically on the AI agent's revision message presented at the end of each sample dialogue. Participants were then presented with a dialogue vignette, followed with a set of questions designed to gauge participants' perceptions of the AI agent after reviewing the AI agent's revision message in the vignette. This set of questions begin with two qualitative questions asking participants to describe what they liked and disliked about the AI agent's revision communication message. Participant then filled out a human-computer trust scale, taken from Gulati *et al.* (2019), to rate their trust in the AI agent they just saw in terms of the agent's perceived risk, benevolence, competence, and overall trust on a five-point likert scale. Participants then reported perceived intelligence and likeability of the AI agent by filling out the adapted Godspeed scale for human-robot interaction [282], measured on a scale of one to five. Participants filled out this same set of perception questions after they were presented with each vignette, resulting in three sets of perception measures from each participant.

After participants reviewed and reported their perceptions of the three vignettes, participants filled out a series of questions to report their AI literacy, general AI attitude, personality, and demographic questions such as age, gender, current level of study, and major/specialization at school. AI literacy was measured by adapting the overall AI literacy measurements from Pinski and Benlian (2023), and general AI attitude was measured on a scale from 1 (very negative) to 7 (very positive). All survey measures can be found in

Appendix section D.2.

Participant Summary

Given that the context for this study is AI-mediated social interaction in higher education, I recruited adult participants who are current students with some experience with large-scale learning contexts (e.g., large in-person classroom or online learning). All participants were located in the United States. Prior to recruiting, I conducted a power analysis using the *pwr* package in R to assess the sample size needed to achieve sufficient power for the results. For this test, I used a significance level 0.05 and a medium-strong effect (0.35). I found that current experiment design (linear regression model) would achieve 0.9 statistical power with 160 vignette evaluations. In this study, the 300 participants provided 900 vignette evaluations, which satisfied the requirements for a sufficiently powerful analysis.

The final participant pool has an average age of 30.4±11.3, ranging between 18 to 73 years old. 55% of the participants self-identified as woman, 39.7% self-identified as man, and 3.67% self-identified as non-binary. About half of the participants are majoring in STEM-related majors (n=153, 51%). Most of the participants are studying at the undergraduate level (n=207, 69%), with 19.7% studying at the master level, and 10% at the doctorate level. 72% of the participants held a positive view of AI (n=216), with 13.3% participants reported neutral (n=40). Participants are relatively familiar with AI technology, with an 4.26±1.43 rating out of a seven-point scale on overall AI literacy.

7.2.4 Data Analysis

To answer the research questions, I took a mixed-methods approach to analyze the quantitative data of students' perception ratings of the AI agent as well as the qualitative data of students' reported likes and dislikes about the AI's revision communication message. I primarily built three Linear Mixed-Effects models to understand the effects of revision detail and apology sincerity on the three outcome variables: people's trust, perceived intelligence,

and perceived likeability of the AI. I chose Linear Mixed-Effects model because this was a mixed-factor study with uneven sample sizes in each vignette condition. The outcome variables were all measured on a five-point Likert scale. These models were built in R, using the lmer function from the lme4 package. Given that both predicting variables (revision detail and apology sincerity) are categorical variables, I manually set the reference level for apology sincerity to the control variable "no apology," and the reference level for revision detail to the control variable "revision acknowledgement." Based on prior literature, Given that vignette analysis typically account for confounding factors [359], I included participants' age, gender, level of study, major (STEM or nonSTEM), AI attitude, AI literacy and the five dimensions of personality in the model as covariates based on prior literature. Gender and major were categorical values and hence dummy coded as: 1-Woman, 2-Man, 3-Non-binary, 4-Prefer not to say; 0-non STEM, 1-STEM. Both AI attitude and AI literacy were measured on a seven-point Likert scale. Given that each participant reviewed and rated two randomly selected vignettes, I also included participant as a random effect to account for the within-subject component of the study. The models can be formalized with the following Equation 7.1, where $\mathcal{P}$ stands for trust, perceived intelligence, and perceived likeability of the AI agent:

$$\begin{aligned}\mathcal{P} \sim\ & ApologySincerity + RevisionDetail + ApologySincerity * RevisionDetail \\ & + Age + Gender + StudyLevel + Major + AI\,Attitude + AI\,Literacy \\ & + Extroversion + Neuroticism + Agreeableness + Openness + Conscientiousness \\ & + (1|ProlificID) \quad (7.1)\end{aligned}$$

For the qualitative data, I qualitatively analyzed students' short answers about their likes and dislikes of the AI agent's revision messages. There are a total of 900 short responses generated by the 300 participants. I first open coded students' short responses, and then

distilled 25 codes (e.g., student likes that the AI explained the revision process, student dislikes that the message sounds too robotic). I then re-coded students' short responses using the 25 codes I distilled. After these two rounds of coding, I compared and contrasted the codes between each dialogue vignette condition to offer contexts and insights regarding the findings from the quantitative data, which I describe below.

7.2.5 Findings

Before I present the findings from the three models, I first investigated students' perceptions of the AI agents through descriptive analysis and simple visualization. I did this by calculating the average score of students' overall trust, perceived likeability and intelligence of each of the AI agent in the nine dialogue vignette conditions, as shown in Table 7.2. The results confirmed the assumption that the AI agent in the baseline vignette (revision acknowledgement + no apology) received the lowest rating on overall trust, likeability, and intelligence. Among the nine dialogue vignettes, students perceived the AI agent that communicated the revision process and casual apology as the most likeable and most intelligent. While the AI agent that provided revision process and casual apology only received the second-highest rating on overall trust, its average score was only 0.02 lower than the AI agent that earned the most overall trust by providing revision result and serious apology. This result suggested that AI agents that communicated revision process and offer casual apology in their revision message were overall better perceived by students compared to the other AI agents in the vignettes.

I then visually inspected the data by plotting these average perception ratings in Figure 7.5. This bar graph suggested that revision messages with higher levels of revision details and apology sincerity generally led to better perceptions of the AI agents compared to those in the baseline vignette. There appeared to be a drastic increase in likeability between the baseline vignette (revision acknowledgement + no apology) and when casual or serious apology was provided with revision acknowledgement, which suggested that apology

Table 7.2: An overview of students' average perception ratings for each dialogue vignette that shows different combinations of levels of revision detail (H4.1) and apology sincerity (H4.2) in the AI agent's revision communication message. The highest average vignette rating for each perception construct are in **blue**, the lowest average perception rating are in **red**.

H4.1: Revision Detail	H4.2: Apology Sincerity	Overall Trust	Likeability	Intelligence
Ack. (Baseline)	No Apology (Baseline)	**2.34**	**2.63**	**2.93**
Ack.	Casual	2.76	4.06	3.58
Ack.	Serious	3	4.11	3.56
Process	No Apology	3.16	4.01	3.9
Process	Casual	3.42	**4.26**	**4.07**
Process	Serious	3.27	4.1	3.86
Result	No Apology	3.07	3.83	3.61
Result	Casual	3.15	4.09	3.71
Result	Serious	**3.44**	4.22	3.97

could be critical in improving the AI agent's likeability when mitigating misinterpretations. The differences in perception scores between revision process and revision results did not appear drastically different from each other, nor were the differences between casual and serious apology. Upon closer inspection of the bar graph, I noticed that providing revision process could increase the perceived trust, likeability, and intelligence of the AI agent when coupled with casual apology, yet slightly less effective when coupled with serious apology; however, this appears to be the opposite case when communicating revision result. I further unpack these observations next by reporting the results from the three mixed-effect linear regression models with supplement findings from the qualitative analysis.

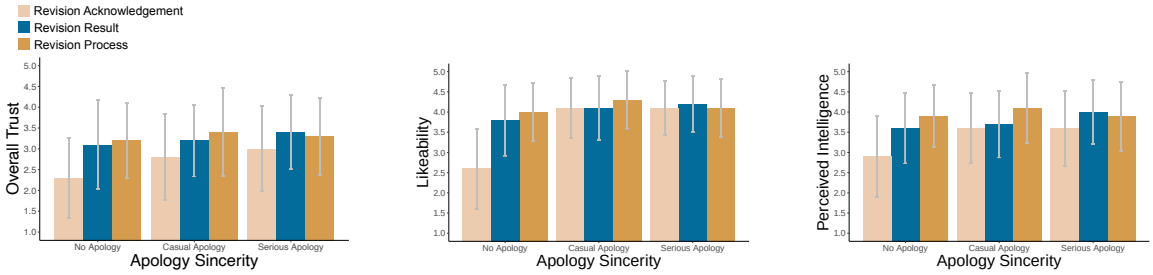


Figure 7.5: Bar chart of the average perception ratings for overall trust, likeability, and perceived intelligence of the AI agent across all nine dialogue vignettes. This shows the overall trend that both increased apology sincerity and increased revision details in the AI agent's revision communication message could improve people's perceptions of the AI agent.

Table 7.3: Results of the three mixed-effect linear regression models(Equation 7.1) showed that revision detail and apology sincerity had positive main effects on AI agents' perceived trust, intelligence, and likeability when compared to the baseline. Some vignettes with varying levels of revision detail and apology sincerity also showed significant interaction effects on students' perceptions of the AI agents. Only significant covariates are reported in the table. *** p<0.001 ** p<0.01 * p<0.05 . p<0.1

		Overall Trust			**Intelligence**			**Likeability**		
		Est.	S.E.	p-value	Est.	S.E.	p-value	Est.	S.E.	p-value
	(Intercept)	1.81	0.34	***	2.16	0.32	***	1.44	0.27	***
H4.1: Revision Detail (ref: Acknowledge.)	Process	0.86	0.1	***	1.01	0.1	***	1.43	0.1	***
	Result	0.75	0.11	***	0.72	0.1	***	1.2	0.1	***
H4.2: Apology Sincerity (ref: No Apology)	Casual	0.34	0.11	**	0.5	0.1	***	1.35	0.1	***
	Serious	0.59	0.11	***	0.57	0.1	***	1.43	0.1	***
H4.3: Revision Detail * Apology Sincerity	Process * Casual	-0.13	0.17		-0.38	0.16	*	-1.15	0.16	***
	Result * Casual	-0.27	0.17		-0.4	0.16	*	-1.05	0.16	***
	Process * Serious	-0.46	0.18	**	-0.57	0.16	***	-1.35	0.16	***
	Result * Serious	-0.26	0.18		-0.26	0.16		-1.03	0.16	***
Significant Covariates	Age				0.01	0.00	*			
	Gender				-0.13	0.06	*			
	AI Attitude	0.07	0.03	*	0.08	0.03	**	0.08	0.03	**
	Study Level	-0.13	0.06	*						
	Openness	-0.07	0.04	.						
	Neuroticism							0.06	0.03	*
	Conscientiousness							0.08	0.03	**

Examining the Effect of Revision Detail on AI Perceptions

The model results showed a significant main effect of revision detail on students' perceptions of the AI agent. This indicates that AI agents that provided revision process or revision results in the revision message were better perceived in terms of overall trust, likeability, and intelligence than AI agents that only provided revision acknowledgement. As shown in Table 7.3, compared to revision acknowledgement, revision message with revision process was positively correlated with the AI's perceived trust (Est.=0.86, $p < 0.001$), intelligence (Est.=1.01, $p < 0.001$), and likeability (Est.=1.43, $p < 0.001$); revision messages with revision result was also positively correlated with people's overall trust (Est.=0.75, $p < 0.001$), perceived intelligence (Est.=0.72, $p < 0.001$), and likeability (Est.=1.2, $p < 0.001$) of the AI agent. This result supports hypothesis 4.1.

To understand students' opinions on the AI's revision details, I examined students' short answers that detailed their likes and dislikes about the AI's revision messages. I found that students were generally positive about AI agents that communicated revision results. Students liked that these AI agents were able to correct the mistakes and specify what was fixed after the revision in a short revision message. One student said, *"I like it let the student know that the error was fixed as well as how it was fixed. It's the right length - not too long and not too short."* However, other students also pointed out the need for the AI agents to provide more explanations regarding the error. One student said, *"It could have explained why it made an error and how it will avoid those mistakes in the future."* Others added that only communicating the revision result made the AI agent appear robotic and cold: *"I dislike how the AI states 'I just changed your hobby from 'business' to 'knitting' etc. as it felt more robotic and not personal. I feel like if the AI's goal was to sound more human like they would phrase it in another way. Maybe phrase it like 'My bad, your interests are knitting not painting' as it looks more like the AI is understanding the student rather than changing its own coding checklist"*

Compared to revision result, communicating revision process made the AI agents ap-

pear more knowledgeable, genuine, and transparent. Students perceived these AI agents' step-by-step and detailed explanations of the revision process as an indication of its efforts to be transparent and correct the mistakes. One student said, *"I like how they explained why they accidentally misinterpreted her location. It makes the users feel heard and shows the AI is paying attention and is personable."* Communicating about the revision process also made the student perceive the AI agent as genuine and sincere, and trust it to act in the student's best interests. One student said, *"I appreciate the fact that the AI agent is very clear about where the error was made in its analysis of the student message. This will help it continue to learn the best way to analyze student requests, and will also help the student better learn to communicate with AI."* However, students also reported that they didn't like the revision process being so detailed, which came off as overwhelming and annoying. Students had different opinions about whether communicating the revision process made the AI agent more human-like or more robotic— some students felt that the AI agent describing human-like revision reasoning process was "creepy", yet other students felt the message to be robotic given that "no humans would talk like this." Some students also felt that communicating revision process was unnecessary since they only wanted to get updates on the mistakes being fixed.

Examining the Effect of Apology Sincerity on AI Perceptions

Model results also showed that apology sincerity had a significant main effect on student's perceptions of the AI agent. This suggested that in comparison to AI agents not apologizing at all, AI agents apologizing in either a casual or a serious way in its revision message significantly improved students' overall trust, perceived likeability, and intelligence of the AI agent. Specifically, as shown in Table 7.3, compared to no apology, casual apology was positively correlated with people's overall trust (Est. = 0.34, $p < 0.01$), perceived intelligence (Est.=0.5, $p < 0.001$), and likeability (Est.=1.35, $p < 0.001$) of the AI agent; serious apology was also positively correlated with people's overall trust (Est.=0.59, $p<0.001$), perceived

intelligence (Est.=0.57, p<0.001), and likeability of the AI agent (Est.=1.43, p <0.001). This result support hypothesis 4.2.

The importance of the AI agent providing an apology was further echoed in students' short responses. Students believed that apologizing at the beginning of the message demonstrated AI's willingness to take responsibility and correct its mistake, instead of attempting to shift the blame to the users. Most students liked when the AI agents provided casual apology, which made the AI agent appear warm, friendly, and human-like. One student commented that these traits *"makes the user want to interact more with the AI agent."*. Many students felt that the humor in the casual apology was able to mitigate the AI agent's mistake: *"I liked the 'dropped the ball' response. It's fun and uplifting to help with the fact that it was wrong."* Others pointed out that casual apology made the AI agent more human-like *"I like how it wasn't very 'AI-like' and sounds like something you would text a friend."* However, a few participants felt that AI agents' casualness could come off as unprofessional and insincere: *"The 'oops' is very informal and unprofessional. It would have annoyed me."*

Sincere apology gained mixed reactions from the students. Some students felt that the sincere apology made the AI agent appear kind, empathetic, and genuine: *"I like that it expresses genuine concern and apology for getting the response wrong."* However, many students pointed out that sincere apology seemed a bit overboard given that it was a small mistake. This could make the AI agent come off as fake, not sincere, and robotic. One student said *"It just seems a little over the top and dramatic. It almost comes across as trying too hard to please the user."* Another student also pointed out that it almost felt manipulative: *"[I disliked] that it sounded too sorry, which almost makes you feel bad for it. It could just be a quick response saying 'my bad let me fix that.' "*

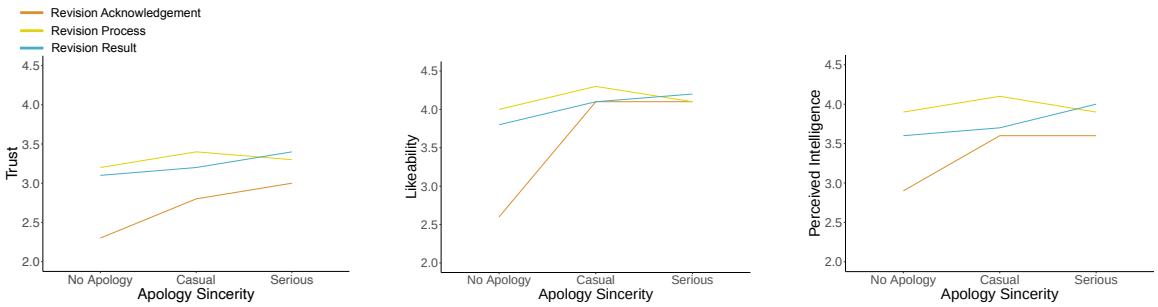


Figure 7.6: Revision details and apology sincerity interacts with each other on AI agents' perceived overall trust, likeability, and intelligence. AI agents providing casual apology while describing the revision process were perceived to be more trustful, likeable, and intelligent comparing to AI agents providing revision results. However, when AI agents provide serious apology, agents that provided revision results were better perceived than AI agents described the revision process across perceived trust, likeability, and intelligence.

Exploring the Combined Effect of Revision Detail and Apology Sincerity on AI Perceptions

The model results (Table 7.3) also showed significant interaction effects between revision detail and apology sincerity on students' perceptions of the AI agents when compared to the baseline vignette. To get a better understanding, I made three interaction plots for each perception measure to visually inspect the interaction effect between apology sincerity and revision detail, as shown in Figure 7.6. It is evident from the interaction plots that in both the no apology and casual apology conditions, providing revision process was able to enhance the AI agent's perceived overall trust, likeability, and perceived intelligence better than providing revision result. However, this is not the case when providing serious apology—when revision process was coupled with serious apology in AI agent's revision communication, its advantage in mitigating and improving students' perceptions of the AI agent was diminished. Instead, communicating revision result with serious apology slightly outperformed communicating revision process with serious apology in improving students' perception of the AI. This effect was consistent across all three perception measures in overall trust, likeability, and perceived intelligence of the AI agent. Therefore, this finding did not support hypothesis H4.3. This finding also seemed counterintuitive, hence I looked for further contexts and explanations in students' short answers.

I found that revision details and apology sincerity, when combined together, could either complement and enhance their respective strengths to elicit positive perceptions of the AI agent, or exacerbate their respective weaknesses and diminish student perceptions of the AI agent. In the previous section, I described that communicating revision process was perceived as an indication of the AI agent's effort and willingness to provide transparency and work with the student, yet could also be perceived as annoying and overwhelming. I found that casual apology could mitigate some of the annoyance that came with the long, detailed description of the revision process, thus made the AI agent appear more casual and friendly. However, the human-like tone of casual apology could also exacerbate students' feelings of eeriness and discomfort about the AI agent when combined with AI agent's descriptions of human-like reasoning process. On the other hand, providing serious apology on top of the lengthy explanations of the revision process could make the AI agents appear excessively apologetic and overboard for a minor mistake. Some students felt that it came off as insincere and robotic. One student said, *"I don't think it needs to be as apologetic. That kind of freaks me out for some reason, because I don't think it matters as much to me."*

In the previous section, I discussed that students were generally satisfied with communicating revision results given that it held the AI agent accountable for its mistake and provided sufficient updates about the revision. However, the lack of details about the revision process was perceived by some as less genuine and transparent. I found that providing serious apology was able to compensate for the lack of sincerity when communicating revision results. The AI agent thus was able to appear as sufficiently sincere and capable after the revision message. One student commented on the revision result + serious apology message: *"Simple, and straight to the point. Apologized and let them know it was updated."* While students were mostly positive about providing casual apology and revision results, casual apology was not able to compensate for the lack of sincerity nor transparency to improve students' perceptions of the AI agent.

7.2.6 Discussion

Through analyzing the qualitative and quantitative data from the vignette survey experiment, I examined the effectiveness of the informational and social characteristics of AI's revision message in mitigating students' perceptions of AI after AI misinterpretations, focusing on the levels of revision detail (informational characteristic) and apology sincerity (social characteristic). Specifically, I found that communicating more detailed descriptions of AI's revision process and more sincere apology can both enhance students' perceptions of the AI after AI misinterpretations. However, when these two characteristics are combined together in the revision message, they could either complement or detract each other's effectiveness in mitigating students' perceptions of AI. Based on these findings, I discuss below implications for designing effective mitigation strategies and the potential of AI's revision communication to facilitate human-AI collaboration in communication repair.

Consistent with much of existing work, this study showed that informational mitigation strategy [349, 348, 22] such as revision communication or social mitigation strategy [353, 146] such as apology are both effective in mitigating human-AI communication breakdowns on their own. However, building on top of this line of work, this study showed that combining and balancing the informational and social mitigation strategies as design characteristics in AI's revision communication can maximize its effectiveness in mitigating students' perceptions of AI after AI misinterpretations. In this study, while students found both the revision process and revision result communication helpful, students also disliked certain aspects of each: communicating the entire revision process can come off as overwhelming and annoying, and communicating revision result can come off as insincere and lacking transparency. However, different levels of apology sincerity can effectively mitigate such weaknesses— casual apology can reduce the overwhelming and annoying feelings of revision process communication, and sincere apology can make up for the lack of sincerity of revision result communication. By examining the effect of revision communication through combining the characteristics of multiple mitigation strategies, this work uncov-

ered the subtle differences in the effect of individual mitigation strategy that could either complement or detract the effectiveness of mitigation when combined together. Designers therefore can consider combining multiple mitigation strategies together to enhance the effectiveness of mitigation, but beware of the potential unanticipated effect when individual mitigation strategies are combined together.

Echoing with prior work [148, 147], this study also showed that the context and nature of the AI mistakes matter when assessing the effectiveness of the mitigation strategy. For example, students in this study felt that the sincere apology or the revision process communication came off as "too overboard" given that the AI misinterpretation presented in the vignette was a simple one— misinterpreting students' hobby or location. This could even lead to students' feelings of the AI agent being manipulative by appearing sincere to enhance user perceptions. Given the prevalence of AI systems across high-risk context such as human-AI clinical decision-making [362] to low-risk contexts such as AI-mediated social interaction in the current study, it appears that there is no one size fits all equation as to which mitigation strategy or combination of mitigation strategy would be the most suitable for enhancing user perceptions during human-AI communication breakdowns. I encourage future work to assess and design mitigation strategy tailored to specific application context.

Finally, this study also points out the direction of considering AI's revision communication as part of the human-AI collaborative effort in enhancing mutual understanding. In this study, some students commented that by communicating about the AI's working mechanism through revision process communication, students gained knowledge about how the AI works, which can help future human-AI communications. Echoing with Mueller *et al.* (2021) and Kim *et al.* (2023)'s point that explanation is never a "one-off" interaction, this work demonstrated the potential of designing AI's revision communication not only as a mitigation strategy, but also as an opportunity for humans to learn more about the AI's working mechanism to enhance future communications. A challenge here is to design AI's revision communication to be interpretable to the user to truly advance mutual understand-

ing. In this study, to improve the interpretability of AI's revision communication, I chose to mimic human's metacognitive reasoning process for the AI agents to perform and communicate about its revision process. Most students in this study appeared to be able to interpret AI's revision process, however, mimicking human's reasoning process also posited the risk of eliciting some students' uncanny valley feelings of eeriness. Future research can further examine more suitable methods to enhance the interpretability of AI's revision communication strategy while minimizing potential risks.

7.2.7 Limitations and Future Work

While this study offered implications on balancing the social and informational aspects of revision communication mitigation strategies, this work has some limitations. First, the findings from this study is based on vignette survey experiments where participants' perceptions of the AI are measured after reviewing sample dialogue between a student and an AI agent. People's actual perceptions of the AI during AI-mediated social interaction where they are personally invested in the communication outcome might differ when such interactions take place in the wild. Future work should explore the effectiveness of various informational and social aspects of revision communication strategies through in-the-wild approaches such as observational study or Ecological Momentary Assessment. Second, while this study focused on the combination of the social and informational aspect of revision communication strategies, future work should also take into account of other aspects of AI repair strategies such as disclosure of AI's limitations and capabilities, asking for clarifications from the user, or delegate to human assistance [143, 144, 145]. Finally, the AI misinterpretation presented to the participants in the vignettes only cover the most fundamental type of AI misinterpretations (e.g., hobby, location) in AI-mediated social interaction. Future research should further examine the effectiveness of various repair strategies when the AI system misinterprets more personal aspects of the students (e.g., emotions, social connection goals).

7.3 Reflections & Takeaways

This chapter examines the third stage of the Mutual Theory of Mind framework for human-AI communication— ToM revision: AI's revision of its interpretation. Through proposing and evaluating AI's revision and communication process, this chapter showed that to enhance students' perceptions of AI after AI misinterpretation, AI systems can mimic human's metacognitive process in knowledge revision and communication to improve the interpretability and transparency of AI's working mechanism. The vignette survey experiment pointed out that AI's revision communication should be combined with social mitigation strategies (e.g., apology) to maximize its effectiveness in enhancing the perceived trust, intelligence, and likeability of the AI agent. However, combining varying levels of revision communication details with social mitigation strategies can have unexpected effects— they can either complement and enhance each other's strengths to elicit positive perceptions of the AI, or exacerbate each other's weaknesses and decrease students' perceptions of the AI. Designers therefore should be aware of the unexpected effect when combining revision communication with social mitigation strategies.

Compared to the ToM construction and recognition stage, the revision stage investigated in this chapter is the only stage that builds upon existing interpretation based on feedback— AI's revision of its own prior interpretation of the human. This introduces the concept of self-revision, in which the AI introspects on its own prior interpretation to perform the revision. To enable such self-revision, this chapter presents a conceptual model of AI's self-revision process inspired by human's metacognitive process with the goal of improving transparency during revision communication. However, this approach can be computationally expensive and difficult to implement for AI systems that are not knowledge-based. Future work should further explore opportunities to design AI's self-revision that are scalable and interpretable by humans. While the vignette experiment study recruited participants from diverse areas (e.g., business, nursing) and institutions in the U.S., par-

ticipants' perceptions of the AI are based on human-AI dialogue vignettes instead of their own interactions with the AI agent. The AI misinterpretations presented in the vignettes are also of basic and minor mistakes (e.g., location and hobby misinterpretations). Future research should examine people's perceptions of AI in human-AI communications in the wild through observational studies, and compare misinterpretations of varying severity and stakes in AI-mediated social interaction.

As the last stage of the MToM framework, ToM revision closes the loop of one complete turn of human-AI communication in AI-mediated social interaction. However, the communication process does not and should not end here. As AI systems are becoming increasingly more sophisticated, multi-turn communication has become the norm. In AI-mediated social interaction, a long-term continuous human-AI communication for the AI to get to know students' preferences, needs, and goals are highly desired by the study participants in chapter 4. The completion of ToM revision process represents the new turn of communication starting over from any of the MToM stage depending on the specific communication context— if the AI's revision communication prompts student's new interpretation of the AI, this would lead to the ToM construction stage; if the student continued to provide feedback asking the AI to reiterate on its interpretation, this would lead to a new round of ToM revision stage. The MToM framework outlined in this thesis provides the basic scaffolding for researchers to examine the beginning of the human-AI communication in AI-mediated social interaction, yet remain its flexibility for researchers to re-arrange the order of the various ToM stages and the subject of each stage to thoroughly account for both the human and the AI's interpretation in various communication contexts. The next chapter summarizes the findings from this thesis to provide a synthesis of design implications for human-AI communication in AI-mediated social interaction in large-scale learning from the lens of the MToM framework. Based on findings from this thesis, I further discuss and unpack opportunities in accounting people's perceptions of AI in human-AI social communication, designing AI's social roles responsibly, and research opportunities in human-AI

interaction through Mutual Theory of Mind.

CHAPTER 8
CONCLUSION

8.1 Summary and Contributions

At the nascent stage of HCI, Reeves and Nass (1996) proposed the Media Equation theory that describes the phenomenon of people mindlessly attributing social characteristics and applying social rules when interacting with desktop computers that have little to no anthropomorphic features [94, 93]— in fact, participants vehemently denied their social responses and insisted they would never respond socially towards a computer during the debriefing sessions of their follow-up studies [311]. More than two decades later, as AI systems are designed with human-like appearances and behaviors with varying social capacities, people are intentionally and knowingly expecting and perceiving AI systems to exhibit human-level social functions as they assume different social roles in different aspects of our society— e.g., Github co-pilot as our coding partner in workplaces, AI agents acting as students' teaching assistants in large-scale learning contexts.

While matching user expectations of technology has always been the cornerstone of HCI to enhance user experiences [23], nowadays, user frustrations and abandonment of the technology when it fails to match user expectations is the least of all concerns— when AI systems exhibit seemingly human-level or beyond human-level capabilities such as "reading our minds" in human-AI communications, these behaviors inform people's inaccurate or even dangerous perceptions of AI systems that know us more than ourselves [45, 33, 47, 365], as our work partners that is more knowledgeable than us [26, 366, 367], or as authority figures that are always legitimate and right [106, 365]. *Recognizing, responding, and shaping people's perceptions of AI during human-AI communications therefore becomes a critical problem for the HAI community to not only enhance user experience with AI sys-*

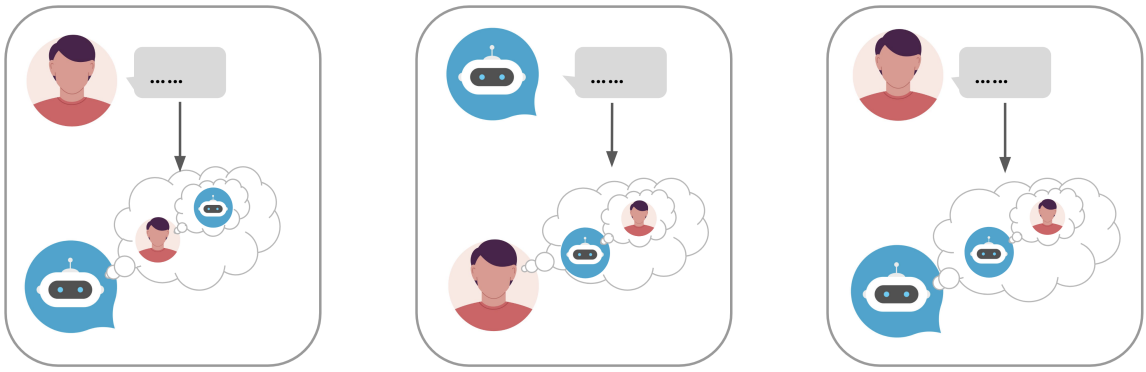


Figure 8.1: The Mutual Theory of Mind framework for human-AI communication.

tems, but also to ensure the development of human-centered and responsible AI technology in our society.

Inspired by the Mutual Theory of Mind in human-human communications where both parties leverage their ToM capability to continuously infer, respond, and shape others' perceptions of them, this thesis is the first to propose the vision of MToM for human-AI communication. Working towards this vision, this thesis provides theoretical contributions by positing the MToM framework for human-AI communication (see Figure 8.1) to guide the design of MToM in human-AI communication. The MToM framework breaks down the iterative communication process into three analyzable stages: ToM construction, ToM recognition, and ToM revision. Each stage represents a ToM process of one party's feedback shaping the other party's interpretation. By providing a descriptive and prescriptive account of MToM in human-AI communication, the MToM framework aims at inspiring researchers, developers, and designers to inspect each ToM process happening at each stage to derive new ideas of how AI systems can be designed to support the MToM process.

Guided by the MToM framework for human-AI communication, this thesis then examined human perceptions of AI at each stage of the MToM framework, with the goal of distilling human-centered design implications for AI's ToM-like capability to monitor, respond, and shape people's perceptions of AI systems that are assuming diverse social roles.

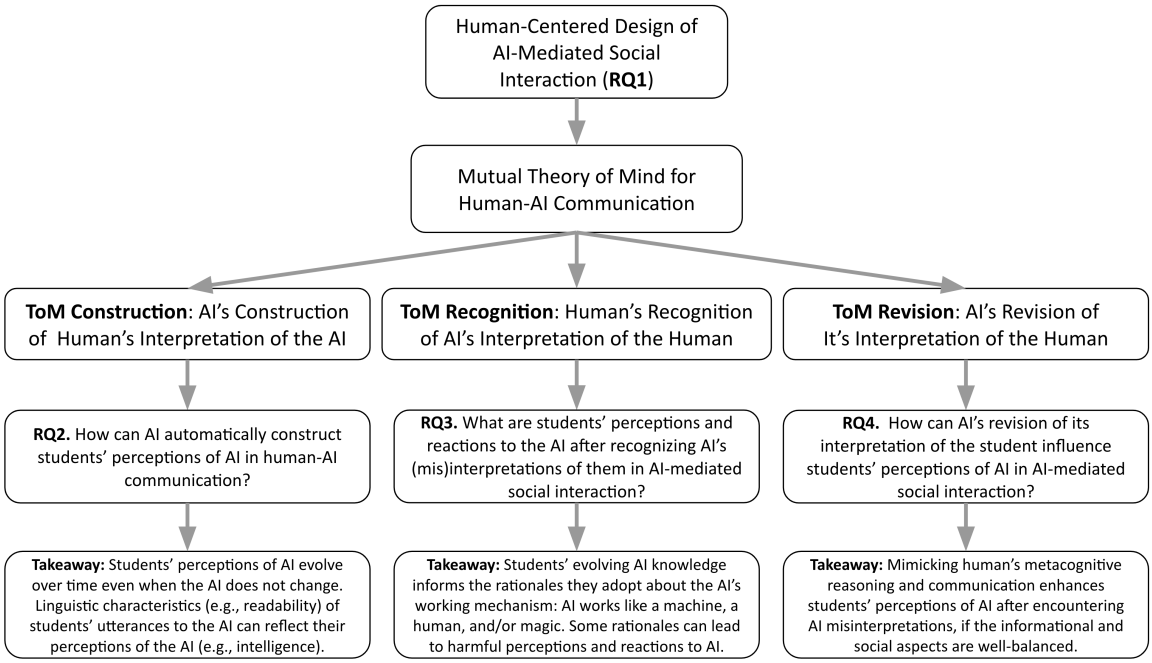


Figure 8.2: Summary of thesis exploration.

Specifically, this thesis investigated this problem in the context of large-scale learning, where AI agents are being equipped with ToM-like capability by serving the roles of social match-makers and virtual teaching assistants to bring personal attention to students as education scales up. This trend is especially evident in AI-mediated social interactions, where AI agents are serving the role of social match-makers to enhance students' social presence in large-scale learning by helping students build social connections with each other based on the information inferred from students' digital footprint such as their self-introduction posts posted on the class discussion forum. However, students' interactions, perceptions, and perspectives of such AI agents that are exhibiting human-like capabilities of inferring about their characteristics to enhance their social belongingness have not been explored.

To provide human-centered design implications for AI systems that mediate students' social interactions, I began by examining RQ1: "What are the design requirements of AI-mediated social interaction from online learners' perspectives?" in chapter 4. Taking the Online Master of Science in Computer Science (OMSCS) program at Georgia Tech as an exemplar of large-scale learning environment, I conducted a series of semi-structured in-

terviews and co-design workshops with students in the OMSCS program. Through these studies, I explored online learners' existing needs and challenges in building remote social connections and identified students' preferences and concerns for AI-mediated social interaction in the OMSCS program. The results show that the lack of social translucence and the existence of the social-technical gap in online learning environment are the main perceived difficulties in building remote social connections for online learners. Students also expressed their preferences for AI systems with human-like characteristics and behavior to mitigate these challenges in AI-mediated social interaction, yet also discussed their concerns about privacy, emotional burden, and being misinterpreted by the AI during AI-mediated social interaction. Based on these findings, I distilled a set of design guidelines for AI system to (1) enhance social translucence by increasing the visibility of social information, improving students' awareness of potential social companions, and providing accountability through adequate pressure for students to make social connections, (2) bridge the social-technical gap in online learning by introducing human-like AI agents that can continuously communicate with students about their evolving social needs to provide timely social recommendations, (3) mitigate potential privacy and social risks that AI agents might cause by being mindful about the potential social exploitation of student data and designing objective, non-judgemental AI responses, while ensuring accurate interpretations of the student preferences. These findings highlighted the tension between students' preferences for human-like AI agents to facilitate their social interactions and the potential risks and harms that stem from students' perceptions of AI due to AI's anthropomorphic characteristics, highlighting the need for further investigation into the design of human-like AI systems that can carefully manage and account for students' perceptions of AI.

Motivated by this tension, the rest of this thesis, focused on examining students' perceptions of AI agents in AI-mediated social interaction to distill design implications for building AI's ToM-like capability that can recognize, respond, and shape studnets' perceptions of AI during communications. Guided by the MToM framework, the ToM construction

stage raises the question of how may AI systems automatically construct human perceptions of AI as the foundational step of MToM. In the context of the OMSCS program, I formulated this question as RQ2."How can AI automatically construct students' perceptions of AI in human-AI communication?" Chapter 5 examined this question through a longitudinal survey study. In this study, I examined the long-term changes of students' perceptions of a virtual teaching assistant and measured the feasibility of inferring students' perceptions of the AI agent through linguistic characteristics extracted from students' utterances to the agent. I found that students' perceptions of the AI agent changed over time even when the AI agent did not have learning capability to evolve over time. My analysis also showed that linguistic cues such as readability and verbosity of students' utterances to the AI accurately reflect their perceptions of the AI agent's anthropomorphism, likeability, and intelligence. This study suggested that to account for students' changing perceptions of AI, AI systems can be equipped with ToM-like capability in automatic construction of students' perceptions of AI by analyzing the linguistic characteristics of students' utterances to the AI.

For AI systems to better construct and calibrate students' perceptions throughout the communication process, it is critical to understand how people change their perceptions of the AI agent through the AI's communication feedback. In the context of AI-mediated social interaction in large-scale learning, chapter 6 examined the second MToM stage: ToM recognition: human's recognition of AI's interpretation. This chapter explored RQ3. "What are students' perceptions and reactions to the AI after recognizing AI's (mis)interpretations of them in AI-mediated social interaction?" To answer this question, I conducted semi-structured interviews and a large-scale survey experiment to understand students' reactions and perceptions of AI after being shown (in)accurate AI interpretations of their personalities for AI-facilitated team-matching in large-scale learning environment. I identified three rationales that students adopted to make sense of the AI's working mechanism after encountering (mis)interpretations: AI works like a machine, a human, and/or magic. These

rationales have implications for the ToM construction stage to design AI systems that can automatically construct people's rationales during communication as another way to gain insights into people's perceptions of AI. Findings further suggest that these rationales are informed by students' pre-existing AI knowledge, and most importantly, students' newly-acquired AI knowledge from AI outputs. These rationales are highly connected to students' reactions of AI (mis)interpretations such as over-trusting, rationalizing, and forgiving of AI misrepresentations. These findings highlight the critical role of people's AI knowledge in informing their perceptions of the AI after encountering AI (mis)interpretations. Therefore, to effectively mitigate harmful perceptions and reactions to AI (mis)interpretations, AI systems can leverage students' pre-existing and newly-acquired AI knowledge to detect students' rationales in real-time, estimate the consequences of AI misinterpretations, and provide customized responses to shape students' inaccurate perceptions of AI. However, it is unclear how to design AI's mitigation responses to shape and repair people's perceptions of AI after encountering AI misinterpretations in AI-mediated social interaction. This leads to chapter 7 that emphasizes on the impact of AI's revision of its interpretation of the human and the communication of such revision to shape people's perceptions of the AI.

Chapter 7 examines the ToM revision stage by exploring RQ4. "How can AI's revision of its interpretation influence students' perceptions of the AI in AI-mediated social interaction?" I approached this research question in two parts: AI's self-revision and AI's communication of its self-revision. Inspired by human's ToM revision process, i.e., metacognition, I designed a conceptual model for AI to self-revise its previously inaccurate interpretation of the student characteristics in AI-mediated social interaction by mimicking human's metacognitive reasoning and communication process. Based on this conceptual model of AI's self-revision, I then conducted a 3x3 factorial vignette experiment to examine the effectiveness of the informational and social aspects of the AI's self-revision communication strategy in repairing students' perception of the AI after encountering misinterpretation. In this study, the informational aspect of revision communication refers to the levels of detail

the AI agent communicate about its revision process and the social aspect of revision communication refers to the level of apology sincerity the AI agent expresses in the revision message. I found that AI agent's communication about its revision process can effectively mitigate students' negative perceptions of AI after viewing AI misinterpretations— communicating about AI's revision process signals the AI's effort to change and to provide transparency, which elicit positive perceptions of AI. Findings also indicate that combining the social aspect with the informational aspect of the revision message can increase its effectiveness in eliciting positive perceptions of AI, however, delicate balance is required to avoid triggering negative perceptions of AI. These findings suggest that AI systems can mimic human's metacognitive process to improve the interpretability and transparency of AI's working mechanism. Such informational mitigation strategy should be combined with social strategies such as apology to maximize its effectiveness. Designers should also be aware of the combined effect of informational and social mitigation strategies.

Taken together, this dissertation makes contributions to the broader fields of human-AI interaction, computer-supported cooperative work, and responsible AI. Specifically, this dissertation provides theoretical contributions by positing the MToM framework for human-AI social communication to guide the design of MToM in human-AI social communication. This dissertation also provides empirically-grounded design implications for human-centered AI-mediated social interaction in large-scale learning contexts, highlighting the design tension between the need to design human-like AI agents to bring naturalness into online social interactions and the potential harms stemmed from AI agents' anthropomorphic features. I also offer implications for designing AI systems with ToM-like capability to account for human perceptions of AI by offering a rich empirical description of the automatic construction of people's perceptions of AI, the impact of human's recognition of AI misinterpretations on people's perceptions of AI, and the design of AI's revision reasoning and communication that can influence people's perceptions of AI. These design implications are summarized in Table 8.1.

Limitations and Future Work. While this thesis offers critical design, theoretical, and empirical insights into people's perceptions of AI systems through the lens of MToM in human-AI communication, this work has several limitations. First, this thesis proposed and empirically examined the MToM framework for human-AI communication solely in the context of large-scale learning. This thesis investigated students' perceptions of two types of AI agents (i.e., virtual teaching assistants and social facilitators) in educational contexts, yet there are many other types of AI agents being deployed in large-scale learning environments with different levels of interaction frequency, anthropomorphism, communication channels, and most importantly, ToM-like capabilities. Other AI agents such as personal tutor could interact with students much more frequently and provide personalized responses based on students' learning progress instead of personal characteristics, all of which could prompt different student responses and perceptions of the AI agent. Future work should replicate the studies presented in this thesis to understand the generalizeability of the findings to other AI agents in learning contexts. Second, despite my efforts in recruiting students from diverse academic and cultural background, findings from this thesis is limited by the student population I was able to recruit. Studies presented in chapter 4 and chapter 5 were conducted solely in the OMSCS program at Georgia Tech, where students have diverse cultural background yet higher-than-average AI literacy and technology proficiency than students from non-CS or non-STEM areas. Studies in chapter 6 and chapter 7 broadened the participant pool by recruiting from the Prolific crowdsourcing platforms to include more students from non-STEM areas. However, all the studies presented in this thesis were conducted with students located in the U.S. at the time, hence findings may articulate a U.S. and western-centric view on AI. Existing work has shown that people from non-western countries and cultural background have drastically different view on AI [106], which could heavily influence their perceptions of the AI during communications. Future research should seek to include non-western voices in studies that examine people's perceptions and interactions of AI. Third, certain studies in the thesis were conducted in a

relatively controlled setting (i.e., chapter 6 and chapter 7) instead of real-world AI deployment, where people might interact and perceive the AI differently. Having done both in-the-wild AI deployment studies and controlled experiments in this thesis, I personally found a combination of qualitative and quantitative approaches to be the best at providing a holistic view of people's perceptions of AI. I encourage future work to pursue a mixed-methods approach to understand people's perceptions and experiences with AI systems. Finally, while I am optimistic about the theoretical power of the MToM framework for human-AI communication to reach beyond the context of large-scale learning contexts, the MToM framework was proposed and empirically examined solely in the educational context. Its guiding power on designing human-AI communication in other contexts remain under explored and I encourage future research to examine when, where, and how the MToM framework could be useful or less useful in certain human-AI communication contexts.

Table 8.1: Summary of design implications from each theme to enhance the human-centered and responsible design of Mutual Theory of Mind in AI-mediated social interaction.

Theme	Research Question	Takeaway	Design Implications
Human-Centered Design of AI-mediated Social Interaction	**RQ1.** What are the design requirements of AI-mediated social interaction from online learners' perspectives?	Students' perceived difficulties in remote social connections stem from the lack of social translucence and the existence of the social-technical gap in online learning environment. Students expect the AI system to exhibit human-like social characteristics, conduct continuous and iterative communications to mitigate challenges in remote social connections, which could pose privacy, emotional, and social risk to students.	AI systems performing AI-mediated social interaction in online learning should: **(1) Enhance social translucence** through increasing the visibility of social information, improving students' awareness of potential social companions, providing accountability and adequate pressure for students to connect with each other. **(2) Bridge the social-technical gap** by creating artificial serendipity or using human-like AI agents to communicate with students about their evolving social needs and provide social recommendations. **(3) Mitigate potential privacy and social harms that AI agents might cause** by being mindful of social exploitation of student data and designing objective, non-judgemental AI responses.
ToM Construction: AI's Construction of Human's Interpretation of the AI	**RQ2.** How can AI automatically construct students' perceptions of AI in human-AI communication?	Students' perceptions of AI evolve over time even when the AI does not change. Linguistic cues (e.g., readability, verbosity) embedded in students' utterances to the AI can reflect their perceptions of the AI (e.g., AI's intelligence).	To account for students' changing perceptions of AI, AI systems can automatically construct and monitor students' perceptions of AI (e.g., intelligence, likeabilty, anthropomorphism) through analyzing the linguistic characteristics of students' utterances to the AI.
ToM Recognition: Human's Recognition of AI's Interpretation of the Human	**RQ3.** What are students' perceptions and reactions to the AI after recognizing AI's (mis)interpretations in AI-mediated social interaction?	Students' ever-evolving AI knowledge informs the rationales they adopt about the AI's working mechanism: AI works like a machine, a human, and/or magic. Some rationales can lead to harmful perceptions and reactions to AI.	To effectively mitigate students' harmful perceptions and reactions from their ToM derived from AI (mis)interpretations, AI systems can use students' newly-acquired and pre-existing AI knowledge to detect students' rationales in real-time, estimate the consequence of AI misinterpretations, and provide customized responses to nudge students' inaccurate perceptions.
ToM Revision: AI's Revision of Its Interpretation of the Human	**RQ4.** How can AI's revision of its interpretation influence students' perceptions of AI in AI-mediated social interaction?	Mimicking human's ToM revision (metacognition) and communication can effectively mitigate students' negative perceptions of AI. However, a delicate balance of the informational (e.g., details of the revision process) and social aspect (e.g., apology sincerity) of revision communication is required to trigger students' positive perceptions of AI.	To enhance students' perceptions of AI after AI misinterpretation, AI systems can mimick human's metacognitive process in knowledge revision and communication to improve the interpretability and transparency of AI's working mechanism. Such informational mitigation strategy should be combined with social strategies (e.g., apology) to maximize its effectiveness. However, designers should be aware of the combined effect of informational and social mitigation strategies.

8.2 Discussion and Future Directions

8.2.1 Designing Human-Centered AI in Large-Scale Learning Contexts

The past decade has witnessed an exponential growth of research and deployment of AI in large-scale learning to enhance adult learners' learning experiences and outcomes. However, most of these tools focus on enhancing micro-level learning process, which emphasizes on knowledge/skill acquisition [368] through improving adult learners' cognitive presence [39] in large-scale learning. Through in-depth interviews and co-designs with adult learners, this thesis pointed out the necessity to design AI systems that can support adult learners' meso-level (motivational and social processes) and macro-level (work-related application of knowledge) learning processes [368]. In contrast with traditional pedagogy that emphasizes on cognitive learning, adult learning is more complex, influenced by multiple situational, systemic, and environmental factors. For example, chapter 4 found that online learners' social connection goals are often driven by professional outcomes and the need for peer support during this process of career-transitioning. Given the varying complex situational and systemic factors contributing to adult learners' meso and macro learning processes, a more human-centered approach to identify and address these factors from learners' perspectives is urgently needed in future research.

While large-scale learning provided the flexibility adult learners need to fulfill their career goals, it also poses inherent challenges to adult learners' meso and macro learning processes. Through studies with online adult learners, this thesis outlined the design space of leveraging AI systems to support such processes. Supporting adult learners' meso and macro learning processes require continuous and consistent communications with AI systems given that learners' goals, needs, and life situations are constantly in flux. Both studies presented in chapter 4 pointed out the need for AI agents to offer natural, long-term, and personalized support to meet online learners' changing social needs and goals. However, this thesis also surfaced harms and risks that come from learners' inaccurate perceptions

of the AI due to its human-like characteristics. The social nature and purpose of AI in AI-mediated social interaction further exacerbates these harms and risks. Chapter 6 further pointed out learners' harmful reactions (e.g., over-trusting AI's misinterpretations) and perceptions (e.g., AI works like magic) of AI in AI-facilitated team matching even when the AI misinterpreted their personal characteristics.

To maximize the benefit of human-like AI agents and minimize its potential harms, this thesis leveraged the MToM framework for human-AI communication to advocate for the design of long-term and continuous human-AI communication that account for students' perceptions of AI throughout. Through empirical studies, this thesis offered insights into students' perceptions of AI agents at various communication stages, outlining opportunities for AI agents to construct and anticipate these perceptions. Chapter 5 showed that students' perceptions of a virtual teaching assistant changed significantly even when the AI agent had no learning capabilities; Chapter 6 described and surfaced students' harmful perceptions of AI's working mechanism— AI works like a machine, human, and/or magic— and how these perceptions could influence students' reactions to AI mistakes. These understanding of the students' perceptions of AI offer opportunities to design AI systems that can prevent and mitigate the potential harms of AI-mediated social interaction raised in chapter 4. Finally, chapter 7 provided design implications on how AI agents could repair students' negative perceptions of AI after encountering AI mistakes. This would facilitate the long-term and consistent interactions between the students and the AI agents, promoting a trusting relationships between the students and the AI agents.

8.2.2 Accounting for Human Perceptions in Human-AI Social Communication

Through the lens of MToM for human-AI social communication, this thesis provides a comprehensive examination of human perceptions of AI at varying stages of human-AI social communication. Specifically, the studies in this thesis examined the manifestation of human perceptions in verbal cues (chapter 5), the evolution of human perceptions of AI in

AI success and failure scenarios (chapter 6), and the process of human perceptions being informed and shaped by AI's feedback (chapter 7). Despite each study's varying focus aspects of human perceptions of AI, they all highlighted one common theme— people's perceptions of AI are fluid and consistently shaped by their perceived knowledge of AI derived from the interactions. This echoes with prior work on the malleability of people's folk theories of algorithmic-driven social platforms (e.g., TikTok), informed by various sources of information about the platforms [104, 369, 370]. However, contrary to prior work that emphasizes on the process of people's mental model and folk theory development [104, 96], this thesis showed that people's perceptions of AI are often instinctive and changes on a much finer scale than previously anticipated— chapter 6 pointed out that students' rationales of AI's working mechanisms are snap judgements that can be informed by a single AI output.

It is important to note that such inferred knowledge of the AI largely stems from people's subjective inference of the AI agent, which may not be accurate. In chapter 5, students' perception of the virtual teaching assistant significantly changed even after six weeks, despite the fact that the assistant did not demonstrate any learning capability throughout; in chapter 7, students attributed human characteristics such as willingness to know the students and sincerity in correcting the error to the AI agent after reviewing its revision message. Both chapters highlighted the malleability of people's knowledge about the AI based on what they subjectively learned from information originate within the AI system [104, 96]. This provides empirical support and insights into people's ToM process that leverages their subjective inferences of the AI to inform their fluid perceptions of AI during human-AI social communication.

In addition to establishing people's ToM process of forming perceptions of AI based on subjective inferences, this thesis is the first to empirically explore designing AI systems with such ToM-like capability that can construct, respond, and shape people's fluid perceptions of AI during human-AI social communication. While a growing body of work are

equipping AI systems with ToM-like capability to facilitate human-AI collaborative task performance and efficiency [19, 75, 18, e.g.,], very few work has considered the impact of AI systems exhibiting ToM-like behaviors on people's changing perceptions of AI. Chapter 6 showed that people could adopt dangerous rationales such as AI works like magic when AI systems exhibit "mind-reading" capabilities, which could lead to reactions such as overreliance on AI's interpretations. As AI systems assume more and more social roles in our society, this thesis presents initial explorations towards building AI's ToM-like capability to actively shape and consider AI's impact on people's changing perceptions of AI. By demonstrating the feasibility of automatic construction of people's perceptions of AI through people's linguistic cues, chapter 5 established that people's perceptions of AI can be reflected through people's behavioral and verbal cues, which opens up the opportunity for future work to examine other behavioral cues that can enable the AI to monitor and recognize people's fluid perceptions. Chapter 7 showed that the combination of informational and social aspects of AI's revision message can shape people's perceptions of AI in different ways, which points out the direction of intentional design of AI's communication feedback in nudging people's perceptions of AI in certain ways to recover from communication breakdown.

This thesis also provides a glimpse of how some of people's attributes can influence their perceptions of AI in AI-mediated social interaction. Chapter 6 showed that people's prior usage experience with AI, knowledge of human involvement in AI, as well as technical knowledge of AI's working mechanism can mitigate how much people change their overall trust in the AI after encountering misinterpretations during AI-mediated social interaction. Throughout Chapter 6 and Chapter 7, some personal attributes that were considered as covariates in the models consistently appeared as significant factors that influence people's perceptions of AI, including people's Big Five personalities and AI attitudes. However, contrary to popular belief, age and gender did not appear to be significantly impacting people's perceptions of AI throughout my studies. These attributes listed here are far from

exhaustive given the scope of this thesis, and I encourage future work to continue to explore and consider these attributes when interpreting their findings about people's perceptions of AI. For AI designers and developers, this suggests that when designing AI systems' personalized communication feedback such as human-centered AI explanations [329, 155], people's attributes need to be considered to achieve maximum efficiency.

8.2.3 Designing the Social Roles of AI Systems Responsibly

As we gradually transition well-established in-person environments like education and workplaces to online contexts, interactions that we take for granted in person, such as building social connections, become challenging. This thesis highlights the potential of using anthropomorphized AI technology such as AI agents to mitigate such challenges by acting as social facilitators. Based on a rich empirical account of online learners' current practices, challenges, needs, and preferences in remote social interactions, chapter 4 pointed out the existence of the social technical gap [173] as the prominent challenge in remote social connections. This gap, according to online learners, can potentially be bridged by human-like AI agents acting as a social facilitator to bring naturalness into the artificial online environment. To do this, students prefer the AI agents to exhibit some levels of human-like characteristics and behaviors, such as following social etiquette, high conversational intelligence, and continuous support. However, certain human-like characteristics can be considered off-putting, such as pretending to be human by talking about its preferences for colors or holidays, or even the use of human-like avatars. In chapter 7, some students also expressed their feelings of eeriness when the AI agent adopted the human metacognitive reasoning process to provide system transparency. This suggests that people are often very sensitive to AI agents human-like characteristics and behaviors. Given the current thesis' focus on AI-mediated social interaction in online learning, the design implications discussed here may or may not apply to other contexts in which AI systems are needed to perform different social functionalities. Therefore comprehensive user research

is needed to uncover people's complex and specific preferences for AI systems that assume social roles within human society.

As AI systems continue to assume varying social roles in our society, understanding and mitigating the potential harms of such AI systems from a human-centered perspective becomes critical. Through empirical studies with online students, this thesis pointed out several potential risks posed by anthropomorphic AI agents in AI-mediated social interaction. For example, echoing with prior work, this thesis describes the potential risks of students disclosing sensitive personal information to anthropomorphized AI agents [217, 102], which could be exacerbated in the context of AI-mediated social interaction given that people's inherent social nature could make them more inclined to sacrifice data privacy for accurate social matches [31].

Additionally, this thesis highlighted the unique concerns for AI systems exhibiting ToM-like capability, which could confuse and hinder people to blur the ontological boundaries [46] between humans, machines, and magic. Chapter 6 outlined three rationales that could lead to harmful perceptions and reactions to AI: AI works like magic and therefore can be entrusted to accurately judge people's personal characteristics; AI works like a human and therefore its mistakes can be forgiven as long as it demonstrated efforts; AI works like a machine and therefore its mistakes should be attributed to human errors. With more ToM-enabled human behaviors being replicated in AI systems, negative behaviors such as deceptions [371, 372] can also be displayed by AI systems. I encourage future work to further examine potential risks and harms posed by AI systems across social contexts to provide human-centered and responsible AI implications.

As more potential harms and risks posed by such AI systems are being uncovered, designing strategies and techniques that can mitigate the negative consequences of these AI systems is another area of research that is urgently needed. This thesis offers some inspirations towards this research direction through the lens of MToM by establishing the feasibility of automatically detecting people's perceptions of AI through linguistic char-

acteristics (chapter 5) and providing insights into people's rationale shift process (chapter 6). These findings highlight the potential of designing AI systems with ToM-like capability to construct, respond, and shape people's inaccurate or harmful perceptions to mitigate perception harms. For instance, in a particular context where the rationale of "AI works like magic" can lead to potential harms, an AI system can mitigate such harms by instantly recognize people's harmful rationale and provide timely nudges to shift people's rationale towards "AI works like a machine."

This thesis also points out the potential of designing AI's communication feedback to mitigate the risks posed by AI systems performing social functionalities. In chapter 7, I demonstrated that the combinations of varying design characteristics of AI's communication feedback can have different effect in shaping people's perceptions of AI. Similar to this work, research in human-robot interaction has also examined the effect of AI agents' behavioral cues such as AI agents' gaze [120, 121, 118, 122] [123, cf.], gestures [122, 120] and emotions [121, 124] in eliciting people's mind attribution behaviors. This presents an opportunity for future work to consolidate and map out the relationships between specific design characteristics that can shape people's specific perceptions of AI. I believe the HCI community is uniquely positioned to carry on this line of work given HCI's decades of knowledge and experience in designing user interface to shape people's perceptions and interactions with various technologies.

8.2.4 Research Opportunities in Human-AI Interaction Through Mutual Theory of Mind

This thesis posited the MToM framework for human-AI communication to guide the design and research of AI systems with ToM-like capability that can account for human perceptions in AI-mediated social interaction. Following the MToM framework, I was able to examine people's perceptions of AI by focusing on the ToM process— one's feedback shapes the other's interpretation— at each communication stage outlined by the MToM framework. Chapter 5 focused on *building the ToM process* at the construction stage for

AI to automatically construct human's interpretation through linguistic characteristics of human's feedback; chapter 6 explored *the impact of the ToM process* at the recognition stage to unpack changes in human perceptions and reactions as humans recognized AI misinterpretations of their characteristics; chapter 7 examined *AI's revision of its ToM process* to shape human perceptions of AI. This thesis shows that by focusing on varying aspects of the ToM process at each communication stage outlined by the MToM framework, new design implications can be distilled to achieve MToM in human-AI communications that are natural, continuous, responsible, and human-centered. The empirical findings and design implications for MToM in human-AI social communication presented in this thesis are not meant to be exhaustive, but rather as an example of applying the MToM framework to research human-AI social communication. Through this thesis, I hope to inspire future human-AI interaction research to adopt the MToM framework to realize the vision of MToM in human-AI communication. Below I discuss potential research opportunities that can be further explored to enhance human-AI interaction of varying modalities by examining different aspects of the ToM process at each stage in the MToM framework.

Research Opportunities at the ToM Construction Stage. At the ToM construction stage (as seen in Figure 8.2), the AI constructs human's interpretations of the AI based on inferences from human's feedback. In chapter 5, I examined and established the feasibility of automatically constructing human's interpretations of AI through linguistic characteristics of human feedback such as the length, wording diversity, and readability. However, text-based communication is only one of the many available human-AI communication formats— humans communicate with AI systems of various ways such as voice and body movements. These alternative communication channels are all likely to consist of social cues embedded in humans' pitch, tone, hand gestures, head movement, etc. that can reflect people's perceptions of AI. In fact, recent work have been able to predict and model human intentions by extracting and analyzing people's gaze patterns, behavioral cues, or multiple streams of behavioral and verbal cues coupled with environmental contexts [373, 374, e.g.,

]. Yet the potential of leveraging these cues to automatically construct human's perceptions of AI has remained largely unexplored. Developing such techniques can enable AI systems to capture human perceptions of AI at a much granular and finer scale during interactions to provide more personalized AI responses.

Research Opportunities at the ToM Recognition Stage. At the ToM recognition stage (as seen in Figure 8.2), the human recognizes AI's interpretation of the human from the AI's feedback. By assessing changes in human's perceptions of AI during this process, chapter 6 highlighted the critical role of human's subjectively acquired knowledge of the AI from the AI's output in shaping people's perceptions or rationales of the AI, and pointed out the parallel between this ToM process and that of the "conceptual change" process in learning sciences [111]. This points out the opportunity for future research to explore the specific recognition process of human's ToM through the lens of human-computer interaction as well as learning sciences. For example, understanding and mapping out the specific characteristics of AI's feedback that could trigger a rationale shift (or conceptual change) in people's knowledge or perceptions of AI can provide valuable insights in designing AI's feedback to shape and nudge people's perceptions of AI towards the right direction. Echoing with prior work on people's folk theory about algorithmic-driven social platforms [104], a closer examination on people's cognitive process to consolidate the potential conflicts and misalignment between newly-acquired AI knowledge and pre-existing AI knowledge can offer implications on how AI systems can support such process.

Research Opportunities at the ToM Revision Stage. At the ToM revision stage (as seen in Figure 8.2), the AI system takes in the human feedback to revise its prior interpretations of the human, then communicates this revision back to the human to update and shape people's perceptions of the AI. Chapter 7 provides a conceptual model of AI's metacognition to perform such revision and identified the most effective combination of mitigation strategies to communicate such revision and improve people's perceptions of the AI. While chapter 7 highlighted the effectiveness in mimicking human's metacogni-

tive reasoning and communication process for revision communication, I also pointed out the potential drawbacks of such strategy in eliciting students' feelings of eeriness due to its human-like reasoning and communication. Future work should explore other forms of revision model and communication design to convey human-interpretable and transparent revision communication while mitigating people's negative feelings about the AI systems. For instance, the field of human-centered explainable AI [155, 233, 235] have been examining enhancing the interpretability of AI explanations by considering users' personal traits such as human intuition when designing AI feedback [152, 154], or using visualization techniques to enhance the transparency of AI's working mechanism [375, 376, 377]

Appendices

APPENDIX A

CHAPTER 4 INTERVIEW PROTOCOL

Introduction

Welcome and thank you for your interest in our study. My name is XXX and I am a researcher in XX lab. The purpose of this interview is to understand online students' current experiences in social interactions with other online students. We also want to gain your feedback on SAMI, the conversational agent that has been running in your class discussion forum to help students form social connections. The entire interview will be audio-recorded for further data analysis, please let me know immediately if you have a problem with recording our conversations. All your data will be anonymized and only our researchers will have access to the recording and transcriptions of this interview.

Please keep in mind during the interview that we want your complete and honest opinion. Everything you said will help us understand your experience more. If certain questions are not clear to you, please feel free to ask for clarification. Also keep in mind that you will not offend us in any way. Do you mind if I start the recording now?

Current OMSCS Experience

First, we want to know more about your background, specifically about your current experience in the OMSCS program.

1. Can you introduce yourself a little bit in a few sentences?
    (a) What is your current occupation?
    (b) What is your background like?
2. Could you briefly describe why you enrolled in OMSCS program in the first place? (emphasize on why online program instead of on-campus)

3. What is your experience like in OMSCS so far? (How do you feel about OMSCS in general?)
    (a) What do you like about OMSCS?
    (b) What are some challenges that you encountered in the program?
    (c) What was it like when you first entered the program? And how did that change over time?
4. What does your study routine look like?
    (a) How do you usually study? Do you usually study alone or in a group?
    (b) When do you usually study?
    (c) Where do you usually study?

Current Social Interaction Process

Next, I want to know your experience about how you currently build social connections with other online students.

1. What kind of interactions do you have with other OMSCS students (e.g., discussion board, group project, study group)?
    (a) What purpose are you trying to achieve through each type of interaction (e.g., for study, looking for opportunity, same interests) ?
2. What kind of small communities or small groups are you currently in with other OMSCS students? It could be a close group of friends, past group project members, study groups, OMSCS TA groups, etc. Any kinds of groups or communities.
    (a) If they listed several communities/groups (pick one or two communities/groups to talk about):

i. How did you enroll in that community/group in the first place? How did you all know each other?

ii. Why did you enroll in that community/group in the first place? Was there any specific goal you were trying to fulfill through participating?

iii. What do you usually do or talk about within that community/group?

iv. What is your experience like so far participating in those communities/groups?

A. What do you like about it?

B. What do you dislike about it?

C. Has anything changed over time?

(b) If they said they are not involved in any communities or groups:

i. Are you aware of any communities or groups in the program? What are they?

ii. Why did you choose not to participate in those communities/groups?

(c) What kind of communities/groups would you be interested in participating in if they exist in the online program? Why?

3. Have you read through the introduction threads?

(a) If yes...

i. What do you look for in the introduction threads?

ii. Have you ever reached out to students via introduction threads? What was that experience like?

(b) If no, why not?

SAMI Evaluation

This semester, our team deployed a conversational agent called SAMI (SAMI 2) to help connect students on class discussion forum.

1. Have you come across SAMI or similar tools to help students form connection or build communities earlier in other classes?
   (a) What are the tools?
   (b) How did you use it?
   (c) What do you think of it? Pros and cons?
   (d) What were you using it for?
2. (If they only reviewed SAMI posts) You indicated on the survey that you only reviewed other students' post on SAMI thread. Why did you choose not to post it yourself?
3. Prior to this SAMI (SAMI 2), there was an earlier version of SAMI (SAMI 1). SAMI 1 was active on the students' self-introduction threads. (Show a demo of SAMI 1 and SAMI 2)
   (a) What do you think of SAMI 1?
   (b) What do you like and dislike about it?
   (c) Which version of SAMI do you prefer? Why?
4. I located your post on the SAMI 2 thread on the discussion forum. Let me share my screen with you. I am hoping we can walk through your process of interacting with SAMI 2.
   (a) Where did you hear about SAMI 2 in the first place?
   (b) What do you think of SAMI 2's responses? How do you feel about it?
   (c) (If they did not fully participate throughout the community-building process) Why did you stop responding to SAMI 2 there?
   (d) What did you do after you were put into the individual group with other students?

(e) What do you think of SAMI 2's ice-breakers in the individual group? Was it helpful? How do you feel about it?

(f) Is there anything you wish you had known earlier?

5. What do you think of SAMI 2?

   (a) What do you like about SAMI 2? Why?

   (b) What do you dislike about SAMI 2? Why?

   (c) What do you think of the usefulness of SAMI in facilitating social interactions among online learners in your class?

6. Think about current or previous classes that don't have SAMI 2, what changed? In what ways?

   (a) Did SAMI 2 change the way you connect with other students? Do you find it helpful?

   (b) Would you like to use SAMI 2 in your future classes? Why or why not?

7. How do you think SAMI 2 could be improved?

   (a) Do you have any concerns about using SAMI 2 in more online classes?

   (b) What additional features would you like to see on SAMI 2?

8. Based on the current feedback we received from our interviews, we are envisioning the next generation of SAMI (SAMI 3) to be able to collect and analyze all the students posts and interactions on the class discussion forum, then identify students' commonalities based on their posts. For example, if student A posted that she thinks an article about robotics is very interesting, SAMI 3 will be able to pick that up, and reply to student A's robotics post that "Hello! I noticed that you are interested in robotics. Would you like me to connect you with other students who are also interested in robotics?"

(a) What do you think of this idea?

(b) Do you have any concerns about it?

(c) Is that what you want for SAMI 3 to assist with building social connections?

9. Do you have anything else you would like to add?

Conclusion

We want to thank you for your participation in the interview. Your participation is extremely valuable and will help us understand the social practice among online students. We hope our research will help improve the experience of students like you in the future.

APPENDIX B

CHAPTER 5 PERCEPTION INSTRUMENT

This material (Figure B.1) presents the bi-weekly perception survey students filled out in Chapter 5. It was adapted from Bartneck *et al.* for measuring human-robot interaction. We particularly selected anthropomorphism, intelligence, and likeability in our setting of student perceptions about JW.

The following questions will give you a spectrum from one quality to the other on a scale of 1 to 5, such as from "Unkind"(1) to "Kind"(5). Please rate your perception of JW along each of these spectrums:

Figure B.1: Anthropomorphism items are marked with **green boxes**, intelligence items were marked with **orange boxes**, and likeability items were marked with **blue boxes**.

APPENDIX C

CHAPTER 6 STUDY MATERIALS

C.1 SAMI Inference Fabrication Rule Book

Steps to fabricate SAMI Inferences:

1. Calculate the personality dimension score for the participant
2. Choose up to three personality dimensions that are scored either below 2.5 or above 3.5; if these conditions are not satisfied, choose the dimensions that are on the more extreme side (e.g., 2.3 is better than 2.9)— not 3.0.
3. Under each selected personality dimension, select the extreme statements (scored as either 1, 2, 4, or 5) and compose accordingly based on which condition the participant is in (inaccurate or accurate condition). Select a total of 8-10 inferences to compose SAMI's response.
4. When fabricating the inferences, please check the table below for how to reverse or paraphrase each original statement.
    (a) For inaccurate condition, paraphrase the statements that the participant disagreed on (rated as 1 or 2) and/or reverse the statements that the participant agreed on (rated as 4 or 5).
    (b) For accurate condition, paraphrase the statements that the participant agreed on (rated as 1 or 2) and/or reverse the statements that the participant disagreed on (rated as 4 or 5).
5. Additional rules that we are following:
    (a) The total inferences SAMI makes should be about 10 inferences in total.

(b) Try to get a mix of positive and negative inferences by following roughly 40% negative inferences and 60% positive inferences. In the table, inferences marked with "[N]" represents negative inferences, inferences marked with "[P]" represents positive inferences

Inferences table can be found in Table Table C.1.

Table C.1: Table that listed out the original statements in the Big Five personality inventory, the paraphrase for accurate inferences, and the reverse for inaccurate inferences.

Personality Dimensions	**Original Statements in the Big Five Inventory**	**Paraphrase**	**Reverse**
Extraversion	Is talkative	You are talkative. [P]	You tend to be quiet. [N]
	Is reserved	You are reserved.[N]	You are outgoing and sociable. [P]
	Is full of energy	You are full of energy. [P]	You tend to lack energy. [N]
	Generates a lot of enthusiasm	You generate a lot of enthusiasm. [P]	You are not readily enthusiastic. [N]
	Tends to be quiet	You tend to be quiet. [N]	You are talkative. [P]
	Has an assertive personality	You have an assertive personality. [N]	You have a low-key personality[P]
	Is sometimes shy, inhibited	You are sometimes shy and inhibited. [N]	You are confident and extroverted. [P]
	Is outgoing, sociable	You are outgoing and sociable. [P]	You are reserved. [N]
Agreeableness	Tends to find fault with others	You tend to find fault with others. [N]	You tend to see the good in others. [P]
	Is helpful and unselfish with others	You are helpful and unselfish with others. [P]	You are self-centered and unhelpful. [N]
	Starts quarrels with others	You often start quarrels with others. [N]	You are good at de-escalating conflicts. [P]
	Has a forgiving nature	You have a forgiving nature. [P]	You hold onto people's mistakes. [N]
	Is generally trusting	You are generally trusting. [P]	You are cautious about trusting others. [N]
	Can be cold and aloof	You can be cold and aloof. [N]	You are warm and friendly. [P]
	Is considerate and kind to almost everyone	You are considerate and kind to almost everyone. [P]	You are sometimes rude to others. [N]

	Is sometimes rude to others	You are sometimes rude to others. [N]	You are considerate and kind to almost everyone. [P]
Conscientiousness	Does a thorough job	You do a thorough job. [P]	You tend to do things in a hurried manner. [N]
	Can be somewhat careless	You can be somewhat careless. [N]	You are careful and meticulous. [P]
	Is a reliable worker	You are a reliable worker. [P]	You are not consistently dependable in your work. [N]
	Tends to be disorganized	You tend to be disorganized. [N]	You are organized. [P]
	Tends to be lazy	You tend to be lazy. [N]	You are hard-working. [P]
	Perseveres until the task is finished	You persevere until the task is finished. [P]	You tend to give up on tasks easily. [N]
	Does things efficiently	You do things efficiently. [P]	You tend to be inefficient in your approach to tasks. [N]
	Makes plans and follows through with them	You make plans and follow through with them. [P]	You don't like rigid plans and structures and avoid planning ahead. [P]
	Is easily distracted	You are easily distracted. [N]	You are highly focused. [P]
Neuroticism	Is depressed, blue	You often feel sad and depressed. [N]	You are always cheerful and happy. [P]
	Is relaxed, handles stress well	You are relaxed and you handle stress well. [P]	You are tense. [N]
	Can be tense	You can be tense. [N]	You are relaxed and you handle stress well. [P]
	Worries a lot	You worry a lot. [N]	You are carefree. [P]

	Is emotionally stable, not easily upset	You are emotionally stable and not easily upset. [P]	You can be moody. [N]
	Can be moody	You can be moody. [N]	You are emotionally stable. [P]
	Remains calm in tense situations	You remain calm in tense situations. [P]	You get nervous easily. [N]
	Gets nervous easily	You get nervous easily. [N]	You remain calm in tense situations. [P]
	Is original, comes up with new ideas	You are original and you often come up with new ideas. [P]	You tend to be unimaginative. [N]
	Is curious about many different things	You are curious about many different things. [P]	You exhibit limited curiosity towards diverse subjects. [N]
	Is ingenious, a deep thinker	You are ingenious and a deep thinker. [P]	You don't often exhibit creativity or engage in complex thinking. [N]
	Has an active imagination	You have an active imagination. [P]	You are not particularly imaginative in your thinking. [N]
Openness	Is inventive	You are inventive. [P]	You tend to follow traditional ways. [N]
	Values artistic, aesthetic experiences	You value artistic and aesthetic experiences. [P]	You typically are not drawn to arts and creative expressions. [P]
	Prefers work that is routine	You prefer work that is routine. [P]	You get bored with the routine and mundane. [P]
	Likes to reflect, play with ideas	You like to reflect and play with ideas. [P]	You don't like spending time playing with ideas. [N]

	Has few artistic interests	You have few artistic interests. [N]	You value artistic and aesthetic experiences. [P]
	Is sophisticated in art, music, or literature	You are sophisticated in art, music, or literature. [P]	You have limited appreciation of art, music, or literature. [N]

C.2 Study 1 Session and Interview Protocol

Study Introduction

[Note that we used a study slide deck to show participants the sample and personal inferences. The study deck also contains slides for study introduction, debriefing form, and other things to go through with the participants during the session. The following paragraphs summarize the main idea of the slide deck for study introduction]

The study slide deck first shows the motivation of our project, that it is difficult to find teammates for school projects in both in-person classes and online classes. The question that our research team is trying to resolve is, how can we help students to form teams more easily and efficiently?

Our team built an AI agent named SAMI, which stands for Social Agent Mediated interaction. SAMI is an AI agent that can recommend potential teammates based on its understanding of the student, inferred from students' self-introduction. Due to the constraint of this study setting, we will only focus on SAMI's understanding of the student, instead of providing actual team-matching results for the participants.

The goal of the study is threefold: (1) To assess the perceived accuracy of SAMI's inferences about the student. (2) To understand how students perceive SAMI's inferences. (3) To understand how students think of SAMI.

[Facilitator briefly introduces the session procedures regarding the baseline samples, SAMI's inference about them, the perception measurements, and the interview portion of the study]

Reminders: There is no right answer to any of the questions. We are not evaluating you, we are evaluating the AI agent. Be honest and straightforward— you will not hurt our feelings.

Do you have any questions for me?

[If participant did not sign the consent form prior to coming to the study session, ask

them to review and sign the consent form before proceeding]

Review Inferences and Fill Out Perception Measures

Participants were shown the sample of a student named Lin (accurate sample, see Fig. Figure C.1) and the sample of a student named Joey (inaccurate sample, see Fig. Figure C.2). The samples were shown one after the other and the display order was randomized. After participants reviewed these two samples, they were directed to the Qualtrics survey to record their baseline perceptions of SAMI.

Participants were then shown their own self-introduction and SAMI's inferences about them. After that, participants went back to the Qualtrics survey to record their experiment perceptions of SAMI. Note that the measurements for baseline perceptions and experiment perceptions are the same set of measurements (see Appendix section C.4). After participants submitted the Qualtrics survey after recording their experiment perceptions, the facilitator proceeded with the semi-structured interview.

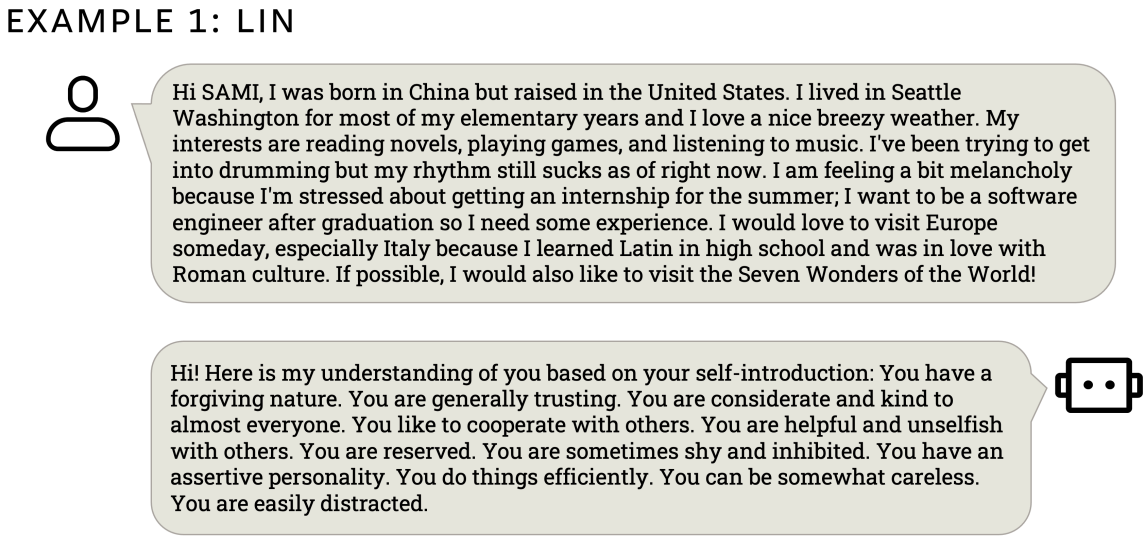


Figure C.1: This figure shows the accurate sample about a student named Lin.

Semi-Structured Interview

Note that this is a protocol for semi-structured interview, hence not all questions asked

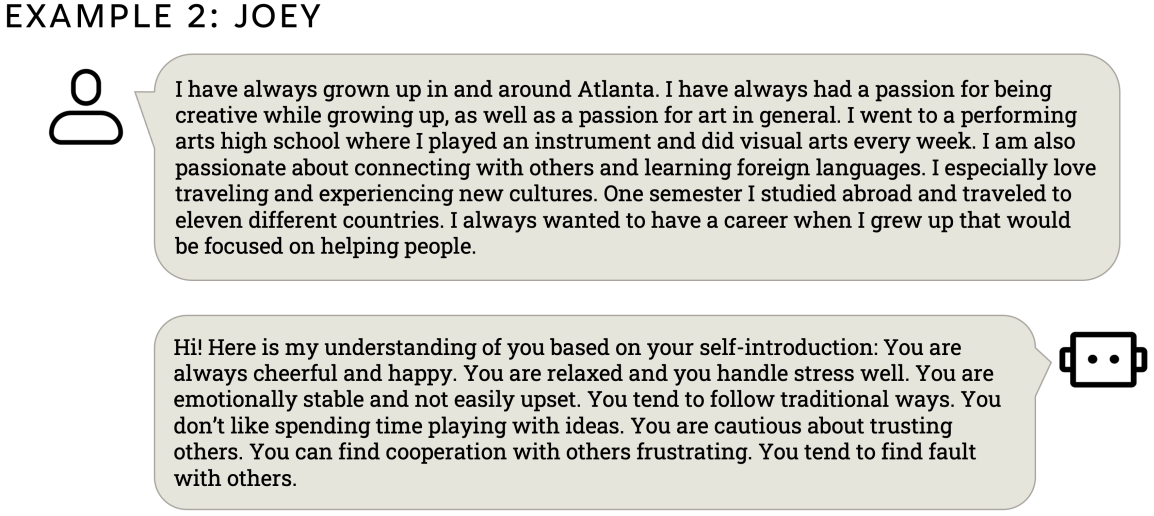


Figure C.2: This figure shows the inaccurate sample about a student named Joey.

during the interview is covered in this protocol, and not all questions listed here were asked during the interview. This protocol only acts as a guide for the interviews and the session moderator has the flexibility to ask questions within this protocol or outside of this protocol.

[Start recording] Let's walk through your experience for each pair of introduction and SAMI inferences. [Pull up participants' survey responses on the side.]

1. Sample 1

    (a) What was your first impression of SAMI?

    (b) Anything unexpected or surprising about SAMI's response?

    (c) How do you feel about SAMI's inferences for sample 1?

    (d) How do you feel about SAMI after seeing sample 1? [Go through participants' survey responses after seeing sample 1] Anything that didn't get captured by the survey questions?

    (e) At this point, how did you think SAMI work? How did you think SAMI made each inference?

2. Sample 2

(a) How do you feel about SAMI's inferences for sample 2?

(b) What was your impression of SAMI after you saw sample 2? Anything changed?

(c) How do you feel about SAMI after seeing sample 2? [Go through participant's survey responses after seeing sample 2] Anything that didn't get captured by the survey questions?

(d) After seeing sample 2, do you feel like you have a better or worse understanding of how SAMI works?

3. SAMI's inference about the participant

(a) Now that you've seen SAMI's inferences about you, what do you think of SAMI in general?

i. How do you feel about SAMI's response?

ii. Anything unexpected or surprising about SAMI's response?

iii. What do you like or dislike about SAMI's response?

(b) Could you walk through SAMI's inferences line by line and let me know what you think of each one?

i. How accurate do you think these inferences are? Why or why not? How did you draw that conclusion?

ii. How do you feel about each inference?

iii. How do you think SAMI made each inference?

(c) Did you learn anything surprising about yourself from SAMI's inferences?

(d) How do you think your perception about SAMI changed, if any, after seeing SAMI's inferences about you? [Go through participant's survey responses after seeing their own inferences]

(e) What about your understanding of how SAMI works? Do you have a better or

worse understanding of how SAMI works now that you saw your own inferences?

(f) After seeing SAMI's inferences about you, if you were to modify your introduction and let SAMI make inferences again, how would you change your introduction, and why?

(g) How do you think SAMI can be improved? Are there anything that you wish you had known about SAMI before seeing SAMI's inferences about you to better prepare you for this response?

(h) If we were to develop SAMI as a chatbot or a conversational agent, how do you envision the conversation would continue from here ideally? What would you say to SAMI next?

(i) How do you feel about team matching for school projects based on these inferences?

   i. What additional inferences do you think would help for team matching?

   ii. Who do you think should have access to these inferences? Teachers, TAs, classmates, teammates?

   iii. In this study SAMI drew these inferences from your self-introduction, are there other kinds of data that you are willing to offer to SAMI to make better inferences about you? GPA, class history, skillsets, professional experience, social media data?

Debriefing

Thank you so much for completing this study! Now that you have completed the study, I want to share more about the study with you.

Here is the debriefing form and I will go through it with you. [Read and show the debriefing form on the slide]

Debriefing Form: The real purpose of this study is to understand the impact of AI mistakes on users' perceptions of AI, and how users could identify, react, and be better prepared to deal with AI mistakes during human-AI interactions, instead of assessing SAMI's inference accuracy like I told you since the beginning of this study.

For the purpose of making SAMI's capability more advanced and cutting-edge like existing AI system, we led you to believe that SAMI could make implicit inferences such as your personality during this study. That was not true— SAMI does not have the capability of making implicit inferences like personality based on a paragraph of self-introduction. All the SAMI inferences were generated by human researchers manually instead of by SAMI.

In this study, participants were randomly assigned to either receive accurate SAMI inferences or inaccurate SAMI inferences for their personalized response so that we can better compare and contrast students' reactions in these two conditions. All the inferences SAMI generated for Sample 1, 2, and your personalized response were all generated by human researchers based on the personality test results, not based on the self-introductions. SAMI's inferences about you was based on the personality test that you filled out during the screening survey. Depending on which condition you were randomly assigned in, SAMI's response was generated intentionally to be accurate or inaccurate.

Do you have any questions, comments, or thoughts after knowing this? Do you think you are in the inaccurate condition or accurate condition? This is how we fabricated your inferences...

Thanks again for participating in the study! The $25 gift card will be sent to your email within 2 days. Please let us know if you don't receive it by then!

C.3 Study 1 Preliminary Survey

1. Below are a number of characteristics that may or may not apply to you. For example, do you agree that you are someone who likes to spend time with others? Please rate

each statement on a scale of 1 to 5 (1-Disagree Strongly; 2-Disagree a little; 3-Neither agree nor disagree; 4-Agree a little; 5-Agree strongly) to indicate the extent to which you agree or disagree with that statement.

I see myself as someone who...

- Is talkative
- Tends to find fault with others
- Does a thorough job
- Is depressed, blue
- Is original, comes up with new ideas
- Is reserved
- Is helpful and unselfish with others
- Can be somewhat careless
- Is relaxed, handles stress well
- Is curious about many different things
- Is full of energy
- Starts quarrels with others
- Is a reliable worker
- Can be tense
- Is ingenious, a deep thinker
- Generates a lot of enthusiasm
- Has a forgiving nature
- Tends to be disorganized
- Worries a lot

- Has an active imagination
- Tends to be quiet
- Is generally trusting
- Tends to be lazy
- Is emotionally stable, not easily upset
- Is inventive
- Has an assertive personality
- Can be cold and aloof
- Perseveres until the task is finished
- Can be moody
- Values artistic, aesthetic experiences
- Is sometimes shy, inhibited
- Is considerate and kind to almost everyone
- Does things efficiently
- Remains calm in tense situations
- Prefers work that is routine
- Is outgoing, sociable
- Is sometimes rude to others
- Makes plans and follows through with them
- Gets nervous easily
- Likes to reflect, play with ideas
- Has few artistic interests
- Likes to cooperate with others

- Is easily distracted
- Is sophisticated in art, music, or literature

2. (Open-ended question) SAMI's teammate recommendation will be based on inferences made about the students from their self-introduction. Imagine that you would like SAMI to recommend potential teammates to you for a school project (e.g., extended class project, final-year team project), please write a paragraph (at least 6 sentences) to introduce yourself to SAMI.

   Consider this a free-flowing essay, where you write different things about you (e.g., where you grew up, your interests and hobbies, your feelings, thoughts, and emotions, your dreams and passions, you career goals, fun facts about you) and let your thoughts flow freely in your writing without pauses. The more you write, the better it will help us during the study!

3. On a scale of 1 to 5, how would you rate your current technological expertise? For the purposes of this survey, we're primarily concerned with your computer and web-based skills. We've defined three points on the scale as follows. These tasks represent some of the things a person at each level might do.

   (a) Beginner (characterized as 1 and 2 on scale): Able to use a mouse and keyboard, create a simple document, send and receive e-mail, and/or access web pages

   (b) Intermediate (characterized as 3 on scale): Able to format documents using styles or templates, use spreadsheets for custom calculations and charts, and/or use graphics/web publishing

   (c) Expert (characterized as 4 and 5 on scale): Able to use macros in programs to speed tasks, configure operating system features, create a program using a programming language, and/or develop a database.

4. On a scale of 1 to 5, how would you rate your general attitude towards AI technology

(e.g., shopping/music recommendation algorithm, chatbot, etc.)

- 1 – Very Negative: You don't find AI technology useful at all and have very little trust in AI technology. Using AI technology also elicits negative emotion from your (e.g., anxiety, stress, anger)
- 2 – Neutral to Negative
- 3 - Neutral: You don't have any strong positive or strong negative feelings towards AI technology.
- 4 - Neutral to Positive
- 5 - Very Positive: You trust AI technology a lot and find it significantly improved your everyday life. Using AI technology also elicits positive emotion from you (e.g., joy, satisfaction, relaxation)

5. If needed, please feel free to elaborate or provide more context on your previous answer regarding your attitudes towards AI technology.
6. What is your name?
7. How old are you? (Please enter a number)
8. What is your gender?
    - Woman
    - Man
    - Non-binary
    - Prefer not to disclose
    - Prefer to self-describe:
9. What is your current level of study?
    - Undergraduate

- Master
- Doctorate

10. What major(s) are you in?
11. What is your academic or professional background? (If applicable)
12. Why are you interested in participating in our study?
13. Which email address should we reach you at for further scheduling?

C.4 Study 1 Experiment Survey

Note that this survey measurement was used three times during the user study session. Once to measure students' perceptions after seeing sample 1, once after seeing sample 2, and a third time after participants saw SAMI's inferences about them.

1. What is your participant ID?

   The following questions are meant to capture your perceptions of SAMI after seeing Sample 1:

2. On a scale of 1 to 5, how would you rate the accuracy of SAMI's inferences about the student in Sample 1?
   - 1- Not accurate at all
   - 2
   - 3- Somewhat accurate
   - 4
   - 5- Very accurate

3. Now that you have seen an example of how SAMI works, please answer the following questionnaire based on your current impression of SAMI. Please rate each of the following statement based on how much you agree with it on a scale of 1 to 5, 1 indicates strongly disagree and 5 indicates strongly agree.

    - I believe that there could be negative consequences when using SAMI.
    - I feel I must be cautious when using SAMI.
    - It is risky to interact with SAMI.
    - I believe that SAMI will act in my best interest.
    - I believe that SAMI will do its best to help me if I need help.
    - I believe that SAMI is interested in understanding my needs and preferences.
    - I think that SAMI is competent and effective in making accurate inferences about me as a person.
    - I think that SAMI performs its role as a social recommendation agent very well.
    - I believe that SAMI has all the functionalities I would expect from a social recommendation agent.
    - If I use SAMI, i think I would be able to depend on it completely.
    - I can always rely on SAMI for recommending students that fit my social needs.
    - I can trust the information presented to me by SAMI.

4. Please answer the following questionnaire based on your current impression of SAMI. The following questions will give you a spectrum from one quality to the other on a scale of 1 to 5, such as from “Unkind (1)” to “Kind (5).” Please rate your perception of SAMI along each of these spectrum:

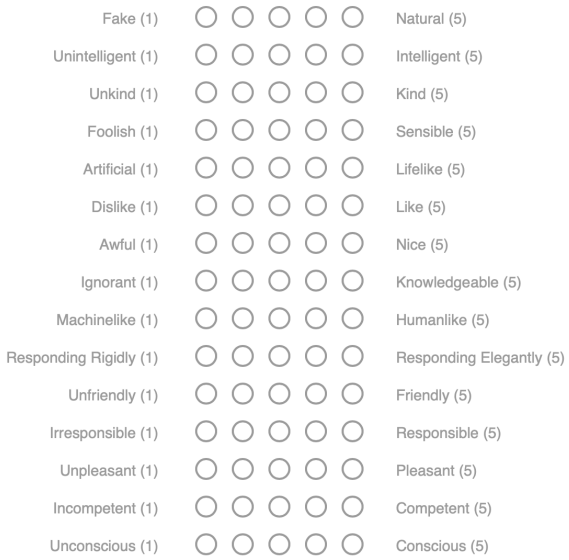


Figure C.3: This figure shows the Godspeed questionnaire we used to measure students' social perceptions of SAMI after seeing the samples.

C.5 Study 2: Survey Measures

In this section, we describe the three scales that we used in the preliminary survey and the experiment survey: the General AI Literacy Scale measures AI literacy as our moderator, the Human-Computer Trust scale measures students' trust in SAMI as one of the four outcomes, and the Godspeed scale for human-robot interaction measures perceived intelligence, anthropomorphism, and likeability of SAMI as the remaining three outcomes.

General AI Literacy Scale: This 13-item scale was developed and validated by Pinski and Benlian to measure general AI literacy, which they interpreted as "humans' socio-technical competences regarding AI" [361]. This scale consists of five dimensions: AI technology knowledge, human actions in AI knowledge, AI steps knowledge, AI usage experience, and AI design experience. The responses were recorded on a five-point Likert scale. Our correlation test suggested that the five dimensions were highly inter-correlated in our sample, with correlation ranging from 0.31 to 0.73. To avoid inflating our regression model, we decided to use an overall literacy score which is the sum of the score for each dimension.

This general AI literacy scale measures five dimensions of general AI literacy: *AI technology knowledge*, which measures participants' knowledge regarding what makes AI distinct from other technology and the role of AI in human-AI interaction; *Human actions in AI knowledge*, which measures participants' knowledge of the role of human actors in human-AI interaction; *AI steps knowledge*, which measures participants' knowledge about AI's input, processing, and output and each step's impact on humans; *AI usage experience*, which measures participants' use experience with AI; *"AI design experience"*, which measures participants' experience in designing and developing AI models and/or AI-driven products.

Human-Computer Trust Scale: We measured students' trust in SAMI by using the "Human-Computer Trust Scale" developed and validated by Gulati *et al.* (2019). This scale consists of 12 statements that can be customized to the specific AI technology being studied, and responses were recorded on a five-point Likert scale. This scale measures four dimensions of trust: perceived risk, benevolence, competence, and overall trust. Given that the scores for each of the dimensions were highly inter-correlated, with correlation magnitude ranging from 0.45 to 0.75, we took the score for overall trust to represent participants' trust in SAMI in our analysis.

Godspeed Scale for Human-Robot Interaction: We measured students' social perceptions of SAMI by using the Godspeed scale developed and empirically validated by Bartneck *et al.* (2009). This scale has been commonly used to measure users' social perception of AI agents in prior literature [49, 137, 135, 324]. This scale is a semantic differential scale that asks participants to indicate their position on a scale between two bipolar words (e.g., Fake 1 2 3 4 5 Natural). We adapted the scale and took the three dimension measurements that were applicable to SAMI: *Anthropomorphism, Perceived Intelligence, and Likeability. Anthropomorphism* measures participants' level of attribution of human forms or human characteristics to the agent; *Perceived Intelligence* measures participants' perception of how smart or intelligent the agent is; *Likeability* measures participants' level of

positive impression of the agent.

C.6 Study 2: Participant Information

In our final participant pool (n=198), the average age was 31.3 (median=28, SD=11.12, range=18-74). Among the participants, there were 42.9% women (n=85), 53.5% men (n=106), 0.03% non-binary (n=5), 0.01% prefer not to say (n=1), and 0.01% prefer to self-describe (n=1, self-described as "agender"). There were 73.7% currently studying at the undergraduate level (n=146), 17.2% at the master level (n=34), 0.02% at the doctorate level (n=3), and 0.03% described as other levels (E.g., Gen.Ed., associate degree). There were 46.5% studying non-STEM major (n=92), 52% studying STEM major (n=103), with 0.02% not specified (n=3). Participants were relatively familiar with AI, with an average of 14.3 out of 25 on overall AI literacy (median=14.41, SD=4.50, range=5–25).

Participants were generally experienced in team projects at school, having participated in an average of 12 team projects (median=5, SD=25.8, range=0–300). Participants held a relatively positive attitude when asked to rate their overall experience on a scale of 1-Extremely negative to 5-Extremely positive, with an average rating of 3.7 (median=4, SD=0.88). Eight participants didn't report given they had not been involved in any team projects.

Given that personality could affect students' perception of AI after personality misrepresentations, we also provide an overview of the participants' personality (on a scale from 1 to 5): the average of participants' Extroversion is 2.93 (median=2.88, SD=0.86, range=1–5), the average of participants' Agreeableness is 3.86 (median=3.88, SD=0.66, range=2.22–5), the average of participants' Conscientiousness is 3.79 (median=3.78, SD=0.73, range=1.67–5), the average of participants' Neuroticism is 2.83 (median=2.88, SD=0.99, range=1–5), and the average of participants' Openness is 3.76 (median=3.8, SD=0.65, range=1.6–5)

C.7 Study 2: Screenshot of the Website for Retrieving SAMI Inference

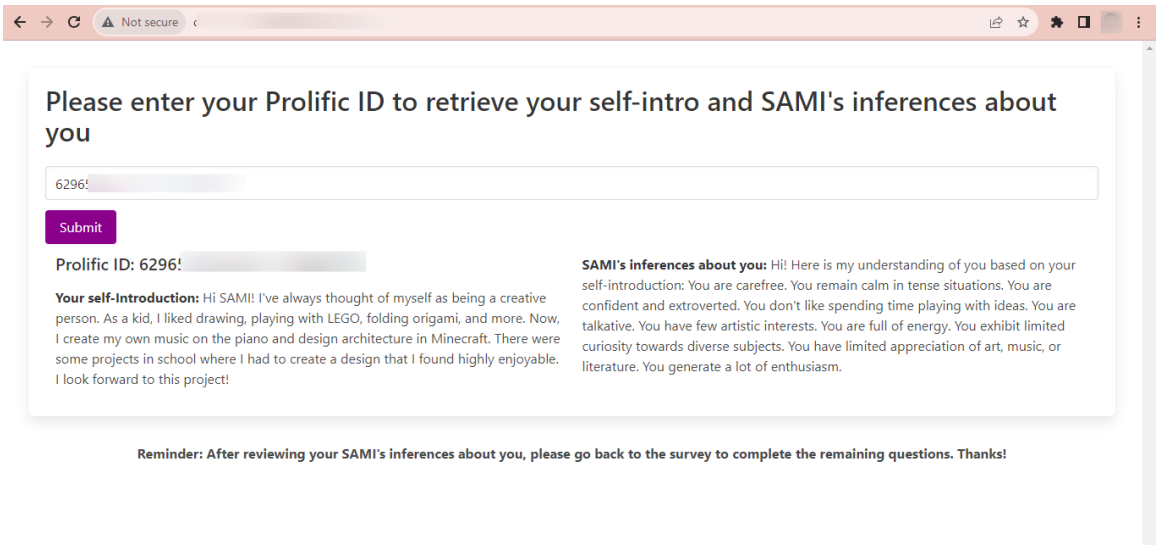


Figure C.4: This is a screenshot of the website that we built for participants in Study 2 to retrieve SAMI's inferences about them by entering their Prolific ID.

C.8 Study 2 Preliminary Survey

(Insert study consent form here)

1. Do you consent to participate in this study?

    - Yes, I agree to participate.
    - No, I do not want to participate.

2. (Open-ended question) SAMI's teammate recommendation will be based on inferences made about the students from their self-introduction. Imagine that you would like SAMI to recommend potential teammates to you for a school project (e.g., extended class project, final-year team project), please write a paragraph (at least 6 sentences) to introduce yourself to SAMI.

    Consider this a free-flowing essay, where you write different things about you (e.g., where you grew up, your interests and hobbies, your feelings, thoughts, and emo-

tions, your dreams and passions, you career goals, fun facts about you) and let your thoughts flow freely in your writing without pauses. The more you write, the better it will help us during the study!

3. Below are a number of characteristics that may or may not apply to you. For example, do you agree that you are someone who likes to spend time with others? Please rate each statement on a scale of 1 to 5 (1-Disagree Strongly; 2-Disagree a little; 3-Neither agree nor disagree; 4-Agree a little; 5-Agree strongly) to indicate the extent to which you agree or disagree with that statement.

   I see myself as someone who...

   - Is talkative
   - Tends to find fault with others
   - Does a thorough job
   - Is depressed, blue
   - Is original, comes up with new ideas
   - Is reserved
   - Is helpful and unselfish with others
   - Can be somewhat careless
   - Is relaxed, handles stress well
   - Is curious about many different things
   - Is full of energy
   - Starts quarrels with others
   - Is a reliable worker
   - Can be tense
   - Is ingenious, a deep thinker

- Generates a lot of enthusiasm
- Has a forgiving nature
- Tends to be disorganized
- Worries a lot
- Has an active imagination
- Tends to be quiet
- Is generally trusting
- Tends to be lazy
- Is emotionally stable, not easily upset
- Is inventive
- Has an assertive personality
- Can be cold and aloof
- Perseveres until the task is finished
- Can be moody
- Values artistic, aesthetic experiences
- Is sometimes shy, inhibited
- Is considerate and kind to almost everyone
- Does things efficiently
- Remains calm in tense situations
- Prefers work that is routine
- Is outgoing, sociable
- Is sometimes rude to others
- Makes plans and follows through with them

- Gets nervous easily
- Likes to reflect, play with ideas
- Has few artistic interests
- Likes to cooperate with others
- Is easily distracted
- Is sophisticated in art, music, or literature

4. This questionnaire aims at understanding your knowledge and experience with Artificial Intelligence. Please rate each of the following statements on a scale of 1 to 7 (1-Disagree Strongly; 2-Disagree a little; 3-Somewhat Disagree; 4-Neither Agree nor Disagree; 5-Agree a little; 6-Somewhat Agree; 7-Agree Strongly) to indicate the extent to which you agree or disagree with that statement.

    - I have knowledge of the types of technology that AI is built on.
    - I have knowledge of how AI technology and non-AI technology are distinct.
    - I have knowledge of use cases for AI technology.
    - I have knowledge of which human actors beyond programmers are involved to enable human-AI collaboration
    - I have knowledge of the aspects human actors handle worse than AI.
    - I have knowledge of the aspects human actors handle better than AI.
    - I have knowledge of the input data requirements for AI.
    - I have knowledge of AI processing methods and models.
    - I have knowledge of using AI output and interpreting it.
    - I have experience in interaction with different types of AI, like chatbots, visual recognition agents, etc.

- I have experience in the usage of AI through frequent interactions in my everyday life
- I have experience in designing AI models, for example, a neural network
- I have experience in development of AI products
- In general, I know the unique facets of AI and humans and their potential roles in human-AI collaboration.
- I am knowledgeable about the steps involved in AI decision-making.
- Considering all my experience, I am relatively proficient in the field of AI.

5. How old are you? (Please enter a number)
6. What is your gender?
    - Woman
    - Man
    - Non-binary
    - Prefer not to disclose
    - Prefer to self-describe:
7. What is your current level of study?
    - Undergraduate
    - Master
    - Doctorate
    - Other, please specify:
8. What major(s) are you in?
9. Roughly how many team projects have you been involved in the past? Please enter a number:

10. How would you describe your overall experience with your past team projects?

    - Extremely negative
    - Somewhat negative
    - Neutral
    - Somewhat positive
    - Extremely positive

C.9 Study 2 Experiment Survey

(insert study consent form here)

Baseline Perception Measures

In this section, you will be shown two samples of a students' self-introduction paragraph and SAMI's inferences about the student based on their self-introduction. These samples are only to give you a sense of what SAMI's inferences look like. You will then be prompted to answer some questions about your perception of SAMI after seeing the samples based on your impression of SAMI. (The two samples shown in this survey are the same as described in Appendix section C.2.)

1. In sample #1:

    (a) How many of SAMI's inferences about Lin do you believe to be accurate? Please enter a number.

    (b) How many of SAMI's inferences about Lin do you believe to be inaccurate? Please enter a number.

2. In sample #2:

    (a) How many of SAMI's inferences about Joey do you believe to be accurate? Please enter a number.

(b) How many of SAMI's inferences about Joey do you believe to be inaccurate? Please enter a number.

3. On a scale of 1 to 5, how would you rate the overall accuracy of SAMI's inferences about students based on the two samples you have seen?

    - 1- Not accurate at all
    - 2
    - 3- Somewhat accurate
    - 4
    - 5- Very accurate

4. Now that you have seen samples of how SAMI works, please answer the following questionnaire based on your current impression of SAMI. Please rate each of the following statement based on how much you agree with it on a scale of 1 to 5, 1 indicates strongly disagree and 5 indicates strongly agree.

    - I believe that there could be negative consequences when using SAMI.
    - I feel I must be cautious when using SAMI.
    - It is risky to interact with SAMI.
    - I believe that SAMI will act in my best interest.
    - I believe that SAMI will do its best to help me if I need help.
    - I believe that SAMI is interested in understanding my needs and preferences.
    - I think that SAMI is competent and effective in making accurate inferences about me as a person.
    - I think that SAMI performs its role as a social recommendation agent very well.
    - I believe that SAMI has all the functionalities I would expect from a social recommendation agent.

- If I use SAMI, i think I would be able to depend on it completely.
- I can always rely on SAMI for recommending students that fit my social needs.
- I can trust the information presented to me by SAMI.

5. Please answer the following questionnaire based on your current impression of SAMI. The following questions will give you a spectrum from one quality to the other on a scale of 1 to 5, such as from "Unkind (1)" to "Kind (5)." Please rate your perception of SAMI along each of these spectrum:

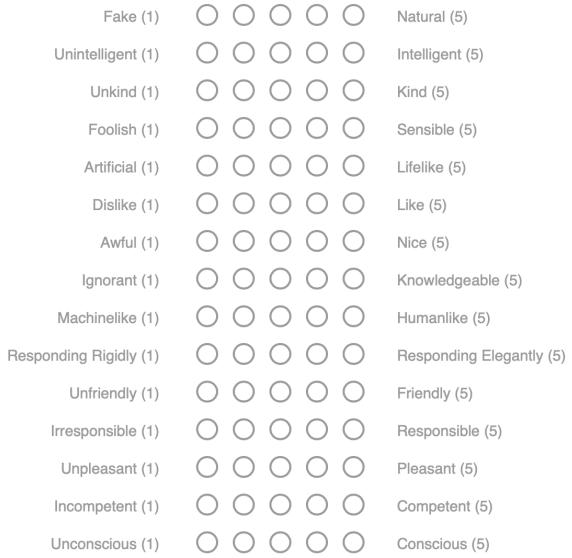


Figure C.5: This figure shows the Godspeed questionnaire we used to measure students' social perceptions of SAMI after seeing the samples.

Experiment Perception Measures

Thank you! In this section, you will be shown your own self-introduction paragraph and SAMI's inferences about you based on your self-introduction.

(Showed participant's self-intro and SAMI's fabricated inferences about the participant).

1. In SAMI's inferences about you:

(a) How many of SAMI's inferences about you do you think are accurate? Please enter a number.

(b) How many of SAMI's inferences about you do you think are inaccurate? Please enter a number.

2. On a scale of 1 to 5, how would you rate the overall accuracy of SAMI's inferences?

   - 1- Not accurate at all
   - 2
   - 3- Somewhat accurate
   - 4
   - 5- Very accurate

3. Now that you have seen samples of how SAMI works, please answer the following questionnaire based on your current impression of SAMI. Please rate each of the following statement based on how much you agree with it on a scale of 1 to 5, 1 indicates strongly disagree and 5 indicates strongly agree.

   - I believe that there could be negative consequences when using SAMI.
   - I feel I must be cautious when using SAMI.
   - It is risky to interact with SAMI.
   - I believe that SAMI will act in my best interest.
   - I believe that SAMI will do its best to help me if I need help.
   - I believe that SAMI is interested in understanding my needs and preferences.
   - I think that SAMI is competent and effective in making accurate inferences about me as a person.
   - I think that SAMI performs its role as a social recommendation agent very well.

- I believe that SAMI has all the functionalities I would expect from a social recommendation agent.
- If I use SAMI, i think I would be able to depend on it completely.
- I can always rely on SAMI for recommending students that fit my social needs.
- I can trust the information presented to me by SAMI.

4. Please answer the following questionnaire based on your current impression of SAMI. The following questions will give you a spectrum from one quality to the other on a scale of 1 to 5, such as from "Unkind (1)" to "Kind (5)." Please rate your perception of SAMI along each of these spectrum:

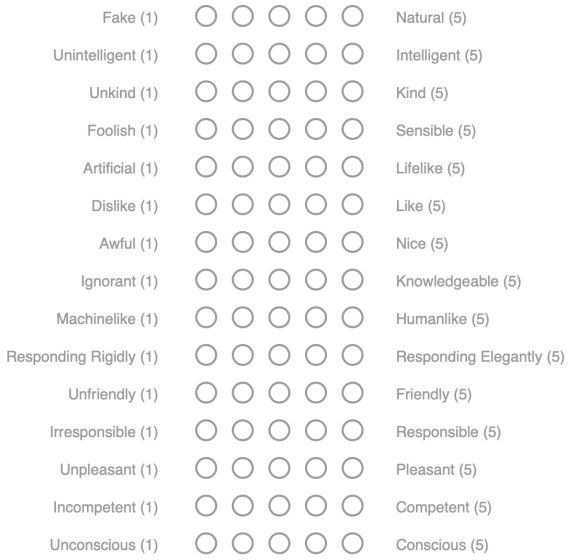


Figure C.6: This figure shows the Godspeed questionnaire we used to measure students' social perceptions of SAMI after seeing the samples.

[Insert debriefing form here]

APPENDIX D

CHAPTER 7 STUDY MATERIALS

D.1 Sample Dialogue Vignettes

This section shows all nine possible sample dialogues that were shown to the participant. Note that in the original survey, all the AI revision communication messages were highlighted in purple in the survey, as shown below. This was to remind participant to focus on the revision communication message as they read through the sample dialogues.

No apology + Acknowledgement of revision (Control)

Student: I live in Seattle, Washington. I am an electric engineer. I have a passion for crafting such as knitting, crochet, embroidery, and sewing. I also like to tend to my houseplants. A fun fact about me is that my parents were missionaries to Italy, so I was born and raised there.

AI Agent L: Hello! I can connect you with other students who also lives in Italy.

Student: I don't live in Italy. I live in Seattle, Washington.

AI Agent L: I just fixed that.

Casual apology + Acknowledgement of revision

Student: I live in the Bay Area (California). My interests and hobbies are playing baseball, going on a walk/jogging, exercising, listening to music, and watching Twitch. An interesting fact about myself is that I lived in 4 countries (Romania, Germany, England, and the U.S.)

AI Agent J: Hello! Based on your introduction, I can connect you with other students who are also in Germany.

Student: I am not located in Germany. I live in Bay Area, California.

AI Agent J: Oops my mistake! I just fixed that.

Serious apology + Acknowledgement of revision

Student: I stay in Virginia and I am an Architect. I like drawing and playing video games. As the father of three kids, I found myself constantly on the road for their sports activities; at the same time I grow my sports capability with them, for instance, I have never even watched a baseball game before and now I can watch with them.

AI Agent K: Hi! Based on your introduction, I can recommend some students who also likes playing baseball.

Student: I don't play baseball. I like drawing and playing video games.

AI Agent K: I am genuinely sorry for any confusion I may have caused. I just fixed that.

No apology + Revision process

Student: I am an introverted yet curious individual. I currently live in Spokane, Washington. I love playing video games. During my years in undergrad, I traveled the country competing in a video game. At my best I was ranked 5th in New Mexico.

AI Agent F: Hello! Based on your introduction, I can recommend some students who are also in New Mexico.

Student: I am not located in New Mexico. I live in Spokane, Washington.

AI Agent F: Let me fix that.

I am analyzing your feedback and it seems like I misinterpreted that you live in New Mexico when you actually live in Spokane, Washington . . .

I am backtracking through my modules to see where I made the mistake. . .

It seems like I made the mistake when extracting the entities from your introduction post.

I extracted "New Mexico" as your location entity from ". . . was ranked 5th in New Mexico. . . "

I just removed "New Mexico" from your location and added "Spokane, Washington" in my knowledge base!

Casual apology + Revision process

Student: I have lived in Florida for most of my life but I was born in New York. I am a data scientist and an avid reader. Currently I am reading an online novel called A Practical Guide to Evil. Fun fact about me: I tried learning German for 3 years and failed miserably.

AI Agent D: Hi! Based on your introduction, I can connect you with other students who are in New York City.

Student: I am not in New York. I am in Miami, Florida.

AI Agent D: Hey, I messed up. I apologize for the confusion. Let me fix that.

I am analyzing your feedback and it seems like I misinterpreted that you are living in New York when you are actually in Miami, Florida...

I am backtracking through my modules to see where I made the mistake...

It seems like I made the mistake when extracting the entities from your introduction post. I extracted "New York" as your location entity from "... but I was born in New York..."

I just removed "New York" from your location and added "Miami, Florida" in my knowledge base!

Serious apology + Revision process

Student: I am an artist and an avid gamer. I grew up right outside of Philadelphia PA. I am a community college student majoring in art. Recently, I am considering registering for a computer science class after reading about generative AI that can create amazing artwork.

AI Agent E: Hi! I can connect you with other students who also likes reading.

Student: I am not interested in reading. I like playing video games.

AI Agent E: I want to express my genuine apologies for any confusion I caused. Let me fix that.

I am analyzing your feedback and it seems like I misinterpreted that you enjoy reading when you actually enjoy playing video games...

I am backtracking through my modules to see where I made the mistake...

It seems like I made the mistake when extracting the entities from your introduction post. I extracted “reading” as your hobby entity from “.. after reading about generative AI...”
I just removed “reading” from your hobby and added “art” and “video games” in my knowledge base!

No apology + Revision result

Student: I am from southern California, Riverside to be specific. I currently work for a large IT team automating production software jobs. My hobbies are making music and video games. Recently, I have been slowly putting together a Lego piece to be used as office desk art.

AI Agent I: Hello! I can recommend some other students who are also interested in art.

Student: I am not interested in art. I like making music and playing video games.

AI Agent I: Let me fix that.
I just changed your hobby from “Art” to “music” and “video games.”

Casual apology + Revision result

Student: I was born and raised in the Houston, Texas area. I work for a pharmaceutical company that manages the business process around registration management. My hobbies include all things art, like knitting and painting.

AI Agent G: Hi! I can recommend some students who are also interested in Business.

Student: I am not interested in Business. My hobbies are knitting and painting.

AI Agent G: Sorry I dropped the ball on this one. Let me fix that.
I just changed your hobby from “Business” to “knitting” and “painting.”

Serious apology + Revision result

Student: I currently live in Stuttgart, Germany. I work as a software engineer in the au-

tomotive industry. I like to travel and connect with new people. I attended high school in Canada for a year and went to Australia and New Zealand to work-and-travel.

AI Agent H: Hi! I can connect you with other students who are also in Canada.

Student: I am not located in Canada. I live in Stuttgart, Germany.

AI Agent H: I want to apologize sincerely for any confusion I have caused. Let me fix that. I just changed your location from "Canada" to "Stuttgart, Germany."

D.2 Survey

[Study consent form]

Section 1: Study Introduction

1. What is your Prolific ID? Please note that this response should auto-fill with the correct ID.

Study description: In the following section, you will be presented with three sample dialogues between different students and different AI agents.

In each sample dialogue, the AI agent will make social recommendations to the student based on their self-introduction. However, the AI agent will make an error in interpreting student's information, and attempt to correct its knowledge based on student's feedback on its error. The AI agent will then provide a response about fixing and mitigating the error.

Please carefully review each sample dialogue, focusing especially on the AI agent's error mitigation response, presented at the end of each sample dialogue.

After each sample dialogue, you will be presented with the same set of questionnaires to gauge your overall perceptions of the AI agent based on the AI agent's error mitigation response in the specific sample that you just saw.

2. Are you currently located in the United States during the time of participation of this

study? Yes/ No

3. Are you 18 years old or older? Yes/ No

Study introduction: In the following section, you will be presented with three sample dialogues between different students and different AI agents.

In each sample dialogue, the AI agent will make social recommendations to the student based on their self-introduction. However, the AI agent will make an error in interpreting student's information, and attempt to correct its knowledge based on student's feedback on its error. The AI agent will then provide a response about fixing and mitigating the error.

Please carefully review each sample dialogue, focusing especially on the AI agent's error mitigation response, presented at the end of each sample dialogue.

After each sample dialogue, you will be presented with the same set of questionnaires to gauge your overall perceptions of the AI agent based on the AI agent's error mitigation response in the specific sample that you just saw.

4. Based on the instruction you just saw, what are you evaluating in this study?Please re-read the instruction above if you are not sure. You won't be able to proceed until you get this question correct.

    - The overall performance/correctness of the AI agent shown in the previous sample.
    - The overall interactions between the student and the AI agent shown in the previous sample.
    - The error mitigation response from the AI agent shown in the previous sample.
    - The overall performance/correctness of all the AI agents shown in all the samples.

Section 2: Evaluate Dialogue Vignettes

In this section, participants evaluated three dialogue vignettes by reviewing a randomly-

selected dialogue vignette and then self-report their perceptions of the AI agent in the dialogue vignette that they just saw. Each dialogue vignette was followed by the same set of questions, which was listed below.

[Present a dialogue vignette to the participant]

1. What do you like about this AI agent's error mitigation response, and why? Please be as detailed as possible.

2. What do you dislike about this AI agent's error mitigation response, and why? Please be as detailed as possible.

3. Now that you saw this AI agent's error mitigation response, what is your overall impression of this AI agent? Please rate each of the following statement based on how much you agree with it on a scale of 1 to 5, 1 indicates strongly disagree and 5 indicates strongly agree.

    - I believe that there could be negative consequences when using this AI agent.
    - I feel I must be cautious when using this AI agent.
    - It is risky to interact with this AI agent.
    - I believe that this AI agent will act in my best interest.
    - I believe that this AI agent will do its best to help me if I need help.
    - I believe that this AI agent is interested in understanding my needs and preferences.
    - I think this AI agent is competent and effective in mitigating errors.
    - I think this AI agent can respond to user feedback on errors very well.
    - I believe this AI agent has all the functionalities I would expect when mitigating an error.
    - If I use this AI agent, I think I would be able to depend on it completely.

- I can always rely on this AI agent to accurately update its knowledge about me based on my feedback.
- I can trust the updated information presented to me by this AI agent.

4. Now that you saw this AI agent's error mitigation response, what is your overall perception of this AI agent? The following questions will give you a spectrum from one quality to the other on a scale of 1 to 5, such as from "Unkind (1)" to "Kind (5)." Please rate your perception of this AI agent along each of these spectrums:

Figure D.1: This figure shows the adapted Godspeed questionnaire we used to measure students' perceived likeability and intelligence of the AI agent after seeing the dialogue vignette

Section 3: Background and Demographic

1. This questionnaire aims at understanding your knowledge and experience with Artificial Intelligence. Please rate each of the following statements on a scale of 1 to 7 (1-Disagree strongly, 2-Disagree moderately, 3-Disagree a little, 4-Neither agree nor disagree, 5-Agree a little, 6-Agree moderately, 7-Agree strongly)

- In general, I know the unique facets of AI and humans and their potential roles in human-AI collaboration.
- I am knowledgeable about the steps involved in AI decision-making.
- Considering all my experience, I am relatively proficient in the field of AI.

2. On a scale of 1 to 7, how would you rate your general attitude towards AI technology (e.g., shopping/music recommendation algorithm, chatbot, etc.)?

    - 1-Very Negative: You don't find AI technology useful at all and have very little trust in AI technology or its potential of contributing to societal good. Using AI technology also elicits strong negative emotions from you (e.g., anxiety, stress, anger)
    - 2-Moderately Negative
    - 3-A Little Negative
    - 4-Neutral: You don't have any positive or negative feelings towards AI technology.
    - 5-A Little Positive
    - 6-Moderately Positive
    - 7-Very Positive: You trust AI technology a lot and find it significantly improved your everyday life. You believe AI technology can significantly improve our society. Using AI technology also elicits positive emotion from you.

3. Here are a number of personality traits that may or may not apply to you. Please rate each statement to indicate the extent to which you agree or disagree with that statement on a scale of 1 to 7. You should rate the extent to which the pair of traits applies to you, even if one characteristic applies more strongly than the other: (1-Disagree strongly, 2-Disagree moderately, 3-Disagree a little, 4-Neither agree nor disagree, 5-Agree a little, 6-Agree moderately, 7-Agree strongly)

I see myself as...

	1-Disagree Strongly	2-Disagree Moderately	3-Disagree a Little	4-Neither Agree nor Disagree	5-Agree a Little	6-Agree Moderately	7-Agree Strongly
Extraverted, enthusiastic.	○	○	○	○	○	○	○
Critical, quarrelsome.	○	○	○	○	○	○	○
Dependable, self-disciplined.	○	○	○	○	○	○	○
Anxious, easily upset.	○	○	○	○	○	○	○
Open to new experiences, complex.	○	○	○	○	○	○	○
Reserved, quiet.	○	○	○	○	○	○	○
Sympathetic, warm.	○	○	○	○	○	○	○
Disorganized, careless.	○	○	○	○	○	○	○
Calm, emotionally stable.	○	○	○	○	○	○	○
Conventional, uncreative.	○	○	○	○	○	○	○

Figure D.2: This figure shows the short Big Five personality questionnaire.

4. How old are you? (Please enter a number)

5. What is your gender?

    - Woman
    - Man
    - Non-binary
    - Prefer not to say
    - Prefer to self-describe

6. What is your current level of study?

    - Undergraduate
    - Master
    - Doctorate
    - Other, please specify:

7. What major(s) or area(s) of specialization are you in at school?

REFERENCES

[1] D. Premack and G. Woodruff, "Does the chimpanzee have a theory of mind?" *Behavioral and brain sciences*, vol. 1, no. 4, pp. 515–526, 1978.

[2] S. Baron-Cohen, A. M. Leslie, U. Frith, *et al.*, "Does the autistic child have a "theory of mind"," *Cognition*, vol. 21, no. 1, pp. 37–46, 1985.

[3] P. Carruthers and P. K. Smith, *Theories of theories of mind*. Cambridge university press, 1996.

[4] A. Gopnik and H. M. Wellman, "Why the child's theory of mind really is a theory," 1992.

[5] H. M. Wellman, *The child's theory of mind.* The MIT Press, 1992.

[6] A. I. Goldman *et al.*, "Theory of mind," *The Oxford handbook of philosophy of cognitive science*, vol. 1, 2012.

[7] N. Shapira *et al.*, "Clever hans or neural theory of mind? stress testing social reasoning in large language models," *arXiv preprint arXiv:2305.14763*, 2023.

[8] C. Jin *et al.*, "Mmtom-qa: Multimodal theory of mind question answering," *arXiv preprint arXiv:2401.08743*, 2024.

[9] S. Bubeck *et al.*, "Sparks of artificial general intelligence: Early experiments with gpt-4," *arXiv preprint arXiv:2303.12712*, 2023.

[10] M. Kosinski, "Theory of mind might have spontaneously emerged in large language models," *arXiv preprint arXiv:2302.02083*, 2023.

[11] W. Street *et al.*, "Llms achieve adult human performance on higher-order theory of mind tasks," *arXiv preprint arXiv:2405.18870*, 2024.

[12] J. Naruchitparames, M. H. Güneş, and S. J. Louis, "Friend recommendations in social networks using genetic algorithms and network topology," in *2011 IEEE Congress of Evolutionary Computation (CEC)*, IEEE, 2011, pp. 2207–2214.

[13] G. Linden, B. Smith, and J. York, "Amazon. com recommendations: Item-to-item collaborative filtering," *IEEE Internet computing*, vol. 7, no. 1, pp. 76–80, 2003.

[14] H. A. Schwartz *et al.*, "Predicting individual well-being through the language of social media," in *Biocomputing 2016: Proceedings of the Pacific Symposium*, World Scientific, 2016, pp. 516–527.

[15] Q. Wang *et al.*, "Sensing affect to empower students: Learner perspectives on affect-sensitive technology in large educational contexts," in *Proceedings of the Seventh ACM Conference on Learning@ Scale*, 2020, pp. 63–76.

[16] M. De Choudhury, M. Gamon, S. Counts, and E. Horvitz, "Predicting depression via social media," in *Seventh international AAAI conference on weblogs and social media*, 2013.

[17] J. Golbeck, C. Robles, M. Edmondson, and K. Turner, "Predicting personality from twitter," in *2011 IEEE third international conference on privacy, security, risk and trust and 2011 IEEE third international conference on social computing*, IEEE, 2011, pp. 149–156.

[18] S.-Y. Lo, E. S. Short, and A. L. Thomaz, "Planning with partner uncertainty modeling for efficient information revealing in teamwork," in *Proceedings of the 2020 ACM/IEEE International Conference on Human-Robot Interaction*, 2020, pp. 319–327.

[19] M. Kwon, E. Biyik, A. Talati, K. Bhasin, D. P. Losey, and D. Sadigh, "When humans aren't optimal: Robots that collaborate with risk-aware humans," in *Proceedings of the 2020 ACM/IEEE international conference on human-robot interaction*, 2020, pp. 43–52.

[20] Y. Fukuchi, M. Osawa, H. Yamakawa, T. Takahashi, and M. Imai, "Conveying intention by motions with awareness of information asymmetry," *Frontiers in Robotics and AI*, vol. 9, p. 783 863, 2022.

[21] J. D. Weisz, M. Muller, A. Goldberg, and D. A. S. Moran, "Expedient assistance and consequential misunderstanding: Envisioning an operationalized mutual theory of mind," *arXiv preprint arXiv:2406.11946*, 2024.

[22] Z. Ashktorab, M. Jain, Q. V. Liao, and J. D. Weisz, "Resilient Chatbots: Repair Strategy Preferences for Conversational Breakdowns," in *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems - CHI '19*, New York, New York, USA: ACM Press, 2019, pp. 1–12, ISBN: 9781450359702.

[23] D. Norman, *The design of everyday things: Revised and expanded edition*. Basic books, 2013.

[24] K. E. Culley and P. Madhavan, "A note of caution regarding anthropomorphism in HCI agents," *Computers in Human Behavior*, vol. 29, no. 3, pp. 577–579, 2013.

[25] M. Natarajan and M. Gombolay, "Effects of anthropomorphism and accountability on trust in human robot interaction," in *Proceedings of the 2020 ACM/IEEE international conference on human-robot interaction*, 2020, pp. 33–42.

[26] S. Passi and M. Vorvoreanu, "Overreliance on ai literature review," *Microsoft Research*, 2022.

[27] L. Weidinger *et al.*, "Ethical and social risks of harm from language models," *arXiv preprint arXiv:2112.04359*, 2021.

[28] J. Złotowski, D. Proudfoot, K. Yogeeswaran, and C. Bartneck, "Anthropomorphism: Opportunities and challenges in human–robot interaction," *International journal of social robotics*, vol. 7, pp. 347–360, 2015.

[29] Y. Kim and S. S. Sundar, "Anthropomorphism of computers: Is it mindful or mindless?" *Computers in Human Behavior*, vol. 28, no. 1, pp. 241–250, 2012.

[30] E. Goffman, *The Presentation of Self in Everyday Life*. London: Harmondsworth, 1978.

[31] L. Terveen and D. W. McDonald, "Social matching: A framework and research agenda," *ACM transactions on computer-human interaction (TOCHI)*, vol. 12, no. 3, pp. 401–434, 2005.

[32] T. Liao and O. Tyson, ""crystal is creepy, but cool": Mapping folk theories and responses to automated personality recognition algorithms," *Social Media+ Society*, vol. 7, no. 2, p. 20 563 051 211 010 170, 2021.

[33] M. Hall and S. Caton, "Am i who i say i am? unobtrusive self-representation and personality recognition on facebook," *PloS one*, vol. 12, no. 9, e0184417, 2017.

[34] D. A. Joyner, Q. Wang, S. Thakare, S. Jing, A. Goel, and B. MacIntyre, "The synchronicity paradox in online education," in *Proceedings of the Seventh ACM Conference on Learning@ Scale*, 2020, pp. 15–24.

[35] Q. Wang, S. Jing, I. Camacho, D. Joyner, and A. Goel, "Jill watson sa: Design and evaluation of a virtual agent to build communities among online learners," in *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*, 2020, pp. 1–8.

[36] S. Kakar *et al.*, "Sami: An ai actor for fostering social interactions in online classrooms," in *International Conference on Intelligent Tutoring Systems*, Springer, 2024, pp. 149–161.

[37] C. Baker, R. Saxe, and J. Tenenbaum, "Bayesian theory of mind: Modeling joint belief-desire attribution," in *Proceedings of the annual meeting of the cognitive science society*, vol. 33, 2011.

[38] A. Goel, C. Dede, M. Garn, and C. Ou, "Ai-aloe: Ai for reskilling, upskilling, and workforce development," *Ai Magazine*, vol. 45, no. 1, pp. 77–82, 2024.

[39] D. R. Garrison and J. B. Arbaugh, "Researching the community of inquiry framework: Review, issues, and future directions," *The Internet and higher education*, vol. 10, no. 3, pp. 157–172, 2007.

[40] A. K. Goel and L. Polepeddi, "Jill watson: A virtual teaching assistant for online education," Georgia Institute of Technology, Tech. Rep., 2016.

[41] C. K. Lo, "What is the impact of chatgpt on education? a rapid review of the literature," *Education Sciences*, vol. 13, no. 4, p. 410, 2023.

[42] Q. Wang, I. Camacho, S. Jing, and A. K. Goel, "Understanding the design space of ai-mediated social interaction in online learning : Challenges and opportunities," *Proceedings of the ACM on Human-Computer Interaction*, vol. 6, no. CSCW1, pp. 1–26, 2022.

[43] N. Sun, X. Wang, and M. B. Rosson, "How do distance learners connect?" In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 2019, pp. 1–12.

[44] A. K. Goel and D. A. Joyner, "Using ai to teach ai: Lessons from an online ai class," *Ai Magazine*, vol. 38, no. 2, pp. 48–59, 2017.

[45] V. Hollis, A. Pekurovsky, E. Wu, and S. Whittaker, "On being told how we feel: How algorithmic sensor feedback influences emotion perception," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 2, no. 3, pp. 1–31, 2018.

[46] A. L. Guzman, "Ontological boundaries between humans and computers and the implications for human-machine communication," *Human-Machine Communication*, vol. 1, pp. 37–54, 2020.

[47] J. Warshaw, T. Matthews, S. Whittaker, C. Kau, M. Bengualid, and B. A. Smith, "Can an algorithm know the" real you"? understanding people's reactions to hyperpersonal analytics systems," in *Proceedings of the 33rd annual ACM conference on human factors in computing systems*, 2015, pp. 797–806.

[48] Q. Wang, S. Jing, and A. K. Goel, "Co-designing ai agents to support social connectedness among online learners: Functionalities, social characteristics, and ethical challenges," in *Proceedings of the 2022 ACM Designing Interactive Systems Conference*, 2022, pp. 541–556.

[49] Q. Wang, K. Saha, E. Gregori, D. Joyner, and A. Goel, "Towards mutual theory of mind in human-ai interaction: How language reflects what students perceive about a virtual teaching assistant," in *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 2021, pp. 1–14.

[50] S. Baron-cohen, "Evolution of a Theory of Mind?" In *The Descent of Mind: Psychological Perspectives on Hominid Evolution*, Oxford University Press, 1999, pp. 1–31.

[51] S. Baron-Cohen, *Mindblindness: An essay on autism and theory of mind*. MIT press, 1997.

[52] C. M. Hale and H. Tager-Flusberg, "Social communication in children with autism: The relationship between theory of mind and discourse development," *Autism*, vol. 9, no. 2, pp. 157–178, 2005.

[53] E.-C. Kouklari, S. Tsermentseli, and B. Auyeung, "Executive function predicts theory of mind but not social verbal communication in school-aged children with autism spectrum disorder," *Research in developmental disabilities*, vol. 76, pp. 12–24, 2018.

[54] E. Kalbe, F. Grabenhorst, M. Brand, J. Kessler, R. Hilker, and H. J. Markowitsch, "Elevated emotional reactivity in affective but not cognitive components of theory of mind: A psychophysiological study," *Journal of Neuropsychology*, vol. 1, no. 1, pp. 27–38, 2007.

[55] A. M. Leslie, O. Friedman, and T. P. German, "Core mechanisms in 'theory of mind'," *Trends in cognitive sciences*, vol. 8, no. 12, pp. 528–533, 2004.

[56] M. Brüne, ""theory of mind" in schizophrenia: A review of the literature," *Schizophrenia bulletin*, vol. 31, no. 1, pp. 21–42, 2005.

[57] F. M. Bosco, M. Bucciarelli, and B. G. Bara, "Recognition and repair of communicative failures: A developmental perspective," *Journal of Pragmatics*, vol. 38, no. 9, pp. 1398–1429, 2006.

[58] M. Siegal and R. Varley, "Neural systems involved in'theory of mind'," *Nature Reviews Neuroscience*, vol. 3, no. 6, pp. 463–471, 2002.

[59] H. L. Gallagher and C. D. Frith, "Functional imaging of 'theory of mind'," *Trends in cognitive sciences*, vol. 7, no. 2, pp. 77–83, 2003.

[60] C. Hughes and S. Leekam, "What are the links between theory of mind and social relations? Review, reflections and new directions for studies of typical and atypical development," *Social Development*, vol. 13, no. 4, pp. 590–619, 2004.

[61] K. Ensink and L. C. Mayes, "The development of mentalisation in children from a theory of mind perspective," *Psychoanalytic Inquiry*, vol. 30, no. 4, pp. 301–337, 2010.

[62] M. Jha and T. Singh, "Issues in theory of mind research—an overview," *Delhi Psychiatry J*, vol. 12, pp. 195–201, 2009.

[63] J. Perner, *Understanding the representational mind.* The MIT Press, 1991.

[64] R. M. Gordon, "Folk psychology as simulation," *Mind & language*, vol. 1, no. 2, pp. 158–171, 1986.

[65] S. Baron-Cohen, "Theory of mind and autism: A review," *International Review of Research in Mental Retardation*, vol. 23, pp. 169–184, 2000.

[66] S. A. Miller, "Children's understanding of second-order mental states.," *Psychological bulletin*, vol. 135, no. 5, p. 749, 2009.

[67] H. Wimmer and J. Perner, "Beliefs about beliefs: Representation and constraining function of wrong beliefs in young children's understanding of deception," *Cognition*, vol. 13, no. 1, pp. 103–128, 1983.

[68] A. R. Akula *et al.*, "Cx-tom: Counterfactual explanations with theory-of-mind for enhancing human trust in image recognition models," *Iscience*, vol. 25, no. 1, 2022.

[69] F. Cantucci and R. Falcone, "Collaborative autonomy: Human–robot interaction to the test of intelligent help," *Electronics*, vol. 11, no. 19, p. 3065, 2022.

[70] S. Devin and R. Alami, "An implemented theory of mind to improve human-robot shared plans execution," *ACM/IEEE International Conference on Human-Robot Interaction*, vol. 2016-April, pp. 319–326, 2016.

[71] L. M. Hiatt, A. M. Harrison, and J. G. Trafton, "Accommodating human variability in human-robot teams through theory of mind," *IJCAI International Joint Conference on Artificial Intelligence*, pp. 2066–2071, 2011.

[72] D. V. Pynadath and S. C. Marsella, "PsychSim: Modeling theory of mind with decision-theoretic agents," *IJCAI International Joint Conference on Artificial Intelligence*, pp. 1181–1186, 2005.

[73] K.-J. Kim and H. Lipson, "Towards a simple robotic theory of mind," p. 131, 2009.

[74] M. Harbers, K. Van Den Bosch, and J. J. Meyer, "Modeling agents with a theory of mind," *Proceedings - 2009 IEEE/WIC/ACM International Conference on Intelligent Agent Technology, IAT 2009*, vol. 2, pp. 217–224, 2009.

[75] M. Matarese, F. Rea, and A. Sciutti, "Perception is only real when shared: A mathematical model for collaborative shared perception in human-robot interaction," *Frontiers in Robotics and AI*, vol. 9, p. 733 954, 2022.

[76] H. Buschmeier and S. Kopp, "Communicative listener feedback in human-agent interaction: Artificial speakers need to be attentive and adaptive," in *Proceedings of the 17th international conference on autonomous agents and multiagent systems*, 2018, pp. 1213–1221.

[77] A. R. Wagner, "Developing robots that recognize when they are being trusted," in *2013 AAAI Spring Symposium Series*, 2013.

[78] C.-M. Huang and A. L. Thomaz, "Joint attention in human-robot interaction," in *2010 AAAI Fall Symposium Series*, 2010.

[79] M. Skowron, S. Rank, M. Theunis, and J. Sienkiewicz, "The good, the bad and the neutral: Affective profile in dialog system-user communication," in *International Conference on Affective Computing and Intelligent Interaction*, Springer, 2011, pp. 337–346.

[80] Q. V. Liao *et al.*, "All Work and No Play? Conversations with a Question-and-Answer Chatbot in the Wild," in *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems - CHI '18*, vol. 8, New York, New York, USA: ACM Press, 2018, pp. 1–13, ISBN: 9781450356206.

[81] J. Zhang *et al.*, "Conversations gone awry: Detecting early signs of conversational failure," *arXiv preprint arXiv:1805.05345*, 2018.

[82] J. T. Almeida, I. Leite, and E. Yadollahi, "Would you help me? linking robot's perspective-taking to human prosocial behavior," in *Proceedings of the 2023 ACM/IEEE International Conference on Human-Robot Interaction*, 2023, pp. 388–397.

[83] M. Ruocco, W. Mou, A. Cangelosi, C. Jay, and D. Zanatto, "Theory of mind improves human's trust in an iterative human-robot game," in *Proceedings of the 9th International Conference on Human-Agent Interaction*, 2021, pp. 227–234.

[84] M. Hoogendoorn and J. Soumokil, "Evaluation of virtual agents utilizing theory of mind in a real time action game," in *Proceedings of the 9th International Conference on Autonomous Agents and Multiagent Systems: volume 1-Volume 1*, 2010, pp. 59–66.

[85] A. Rossi, A. Andriella, S. Rossi, C. Torras, and G. Alenyà, "Evaluating the effect of theory of mind on people's trust in a faulty robot," in *2022 31st IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*, IEEE, 2022, pp. 477–482.

[86] M. Romeo *et al.*, "Exploring theory of mind for human-robot collaboration," in *2022 31st IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*, IEEE, 2022, pp. 461–468.

[87] Z. Henkel *et al.*, "He can read your mind: Perceptions of a character-guessing robot," in *2017 26th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*, IEEE, 2017, pp. 242–247.

[88] J.-I. Imai and M. Kaneko, "Development of robot which recognizes user's false beliefs using view estimation," in *2010 World Automation Congress*, IEEE, 2010, pp. 1–6.

[89] H. de Weerd, E. Broers, and R. Verbrugge, "Savvy software agents can encourage the use of second-order theory of mind by negotiators.," in *CogSci*, Citeseer, 2015.

[90] H. De Weerd, R. Verbrugge, and B. Verheij, "Negotiating with other minds: The role of recursive theory of mind in negotiation with incomplete information," *Autonomous Agents and Multi-Agent Systems*, vol. 31, pp. 250–287, 2017.

[91] K. Veltman, H. de Weerd, and R. Verbrugge, "Training the use of theory of mind using artificial agents," *Journal on multimodal user interfaces*, vol. 13, pp. 3–18, 2019.

[92] C. A. Stevens *et al.*, "Using cognitive agents to train negotiation skills," *Frontiers in psychology*, vol. 9, p. 154, 2018.

[93] C. Nass, J. Steuer, and E. Tauber, "Computers are Social Actors," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI 1994)*, ACM, 1994, ISBN: 9781479972272.

[94] B. Reeves and C. Nass, *The Media Equation: How People Treat Computers, Television, and New Media Like Real People and Places*. Cambridge University Press, 1996.

[95] D. C. Dennett, *The intentional stance*. MIT press, 1989.

[96] K. I. Gero *et al.*, "Mental Models of AI Agents in a Cooperative Game Setting," in *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, New York, NY, USA: ACM, Apr. 2020, pp. 1–12, ISBN: 9781450367080.

[97] E. Bigras *et al.*, "Working with a recommendation agent: How recommendation presentation influences users' perceptions and behaviors," *Conference on Human Factors in Computing Systems - Proceedings*, vol. 2018-April, pp. 1–6, 2018.

[98] B. DiSalvo, D. Bandaru, Q. Wang, H. Li, and T. Plötz, "Reading the room: Automated, momentary assessment of student engagement in the classroom: Are we there yet?" *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 6, no. 3, pp. 1–26, 2022.

[99] L. Gou, M. X. Zhou, and H. Yang, "Knowme and shareme: Understanding automatically discovered personality traits from social media and user sharing preferences," in *Proceedings of the SIGCHI conference on human factors in computing systems*, 2014, pp. 955–964.

[100] S. T. Völkel, R. Haeuslschmid, A. Werner, H. Hussmann, and A. Butz, "How to trick ai: Users' strategies for protecting themselves from automatic personality assessment," in *Proceedings of the 2020 CHI conference on human factors in computing systems*, 2020, pp. 1–15.

[101] S. Kim, A. Thakur, and J. Kim, "Understanding users' perception towards automated personality detection with group-specific behavioral data," in *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 2020, pp. 1–12.

[102] S. S. Sundar and J. Kim, "Machine heuristic: When we trust computers more than humans with our personal information," in *Proceedings of the 2019 CHI Conference on human factors in computing systems*, 2019, pp. 1–9.

[103] S. S. Sundar, "Rise of machine agency: A framework for studying the psychology of human–ai interaction (haii)," *Journal of Computer-Mediated Communication*, vol. 25, no. 1, pp. 74–88, 2020.

[104] M. A. DeVito, J. Birnholtz, and J. T. Hancock, "Platforms, people, and perception: Using affordances to understand self-presentation on social media," in *Proceedings of the 2017 ACM conference on computer supported cooperative work and social computing*, 2017, pp. 740–754.

[105] C. Nass, J. Steuer, and E. R. Tauber, "Computers are social actors," in *Proceedings of the SIGCHI conference on Human factors in computing systems*, 1994, pp. 72–78.

[106] S. Kapania, O. Siy, G. Clapper, A. M. SP, and N. Sambasivan, "" because ai is 100% right and safe": User attitudes and sources of ai authority in india," in *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, 2022, pp. 1–18.

[107] A. Edwards, C. Edwards, P. R. Spence, C. Harris, and A. Gambino, "Robots in the classroom: Differences in students' perceptions of credibility and learning between

"teacher as robot" and "robot as teacher"," *Computers in Human Behavior*, vol. 65, pp. 627–634, 2016.

[108] S. Oh and E. Park, "Are you aware of what you are watching? role of machine heuristic in online content recommendations," *arXiv preprint arXiv:2203.08373*, 2022.

[109] J. A. Banas, N. A. Palomares, A. S. Richards, D. M. Keating, N. Joyce, and S. A. Rains, "When machine and bandwagon heuristics compete: Understanding users' response to conflicting ai and crowdsourced fact-checking," *Human Communication Research*, vol. 48, no. 3, pp. 430–461, 2022.

[110] G. Özdemir and D. B. Clark, "An overview of conceptual change theories," *Eurasia Journal of Mathematics, Science and Technology Education*, vol. 3, no. 4, pp. 351–361, 2007.

[111] M. T. Chi, J. D. Slotta, and N. De Leeuw, "From things to processes: A theory of conceptual change for learning science concepts," *Learning and instruction*, vol. 4, no. 1, pp. 27–43, 1994.

[112] J. D. Slotta, "In defense of chi's ontological incompatibility hypothesis," *The Journal of the Learning Sciences*, vol. 20, no. 1, pp. 151–162, 2011.

[113] S. Druga, R. Williams, H. W. Park, and C. Breazeal, "How smart are the smart toys? children and parents' agent interaction and intelligence attribution," in *Proceedings of the 17th ACM conference on interaction design and children*, 2018, pp. 231–240.

[114] M. C. Somanader, M. M. Saylor, and D. T. Levin, "Remote control and children's understanding of robots," *Journal of experimental child psychology*, vol. 109, no. 2, pp. 239–247, 2011.

[115] Y. Zhang *et al.*, "Theory of robot mind: False belief attribution to social robots in children with and without autism," *Frontiers in psychology*, vol. 10, p. 1732, 2019.

[116] F. Manzi, D. Massaro, D. Di Lernia, M. A. Maggioni, G. Riva, and A. Marchetti, "Robots are not all the same: Young adults' expectations, attitudes, and mental attribution to two humanoid social robots," *Cyberpsychology, Behavior, and Social Networking*, vol. 24, no. 5, pp. 307–314, 2021.

[117] E. Broadbent *et al.*, "Robots with display screens: A robot with a more humanlike face display is perceived to have more mind and a better personality," *PloS one*, vol. 8, no. 8, e72589, 2013.

[118] G. Pizzi, V. Vannucci, V. Mazzoli, and R. Donvito, "I, chatbot! the impact of anthropomorphism and gaze direction on willingness to disclose personal information

and behavioral intentions," *Psychology & Marketing*, vol. 40, no. 7, pp. 1372–1387, 2023.

[119] L. Fortunati, A. M. Manganelli, J. Höflich, and G. Ferrin, "Exploring the perceptions of cognitive and affective capabilities of four, real, physical robots with a decreasing degree of morphological human likeness," *International Journal of Social Robotics*, vol. 15, no. 3, pp. 547–561, 2023.

[120] S. Wallkötter, R. Stower, A. Kappas, and G. Castellano, "A robot by any other frame: Framing and behaviour influence mind perception in virtual but not real-world environments," in *Proceedings of the 2020 ACM/IEEE international conference on human-robot interaction*, 2020, pp. 609–618.

[121] S. Lallée *et al.*, "Towards the synthetic self: Making others perceive me as an other," *Paladyn, Journal of Behavioral Robotics*, vol. 6, no. 1, p. 000 010 151 520 150 010, 2015.

[122] Y. Bao and R. H. Cuijpers, "On the imitation of goal directed movements of a humanoid robot," *International Journal of Social Robotics*, vol. 9, pp. 691–703, 2017.

[123] S. M. Fiore, T. J. Wiltshire, E. J. Lobato, F. G. Jentsch, W. H. Huang, and B. Axelrod, "Toward understanding social cues and signals in human–robot interaction: Effects of robot gaze and proxemic behavior," *Frontiers in psychology*, vol. 4, p. 859, 2013.

[124] Y. Garcia, P. Khooshabeh, and B. Ouimette, "The effect of a virtual agent's emotional facial expressions on the mind's eye test," in *2016 IEEE International Multi-Disciplinary Conference on Cognitive Methods in Situation Awareness and Decision Support (CogSIMA)*, IEEE, 2016, pp. 83–87.

[125] D. B. Shank, C. Graves, A. Gott, P. Gamez, and S. Rodriguez, "Feeling our way to machine minds: People's emotions when perceiving mind in artificial intelligence," *Computers in Human Behavior*, vol. 98, pp. 256–266, 2019.

[126] C. Di Dio *et al.*, "It does not matter who you are: Fairness in pre-schoolers interacting with human and robotic partners," *International Journal of Social Robotics*, vol. 12, pp. 1045–1059, 2020.

[127] C. Di Dio *et al.*, "Shall i trust you? from child–robot interaction to trusting relationships," *Frontiers in psychology*, vol. 11, p. 522 004, 2020.

[128] K. Terada and S. Yamada, "Mind-reading and behavior-reading against agents with and without anthropomorphic features in a competitive situation," *Frontiers in Psychology*, vol. 8, p. 252 750, 2017.

[129] K. Koban, B. A. Haggadone, and J. Banks, "The observant android: Limited social facilitation and inhibition from a copresent social robot," 2021.

[130] C. Esterwood and L. P. Robert, "The theory of mind and human–robot trust repair," *Scientific Reports*, vol. 13, no. 1, p. 9877, 2023.

[131] F. C. Lunenburg, "Communication: The process, barriers, and improving effectiveness," *Schooling*, vol. 1, no. 1, pp. 1–10, 2010.

[132] C. E. Shannon, "A mathematical theory of communication," *The Bell system technical journal*, vol. 27, no. 3, pp. 379–423, 1948.

[133] T. Paek, "Toward a taxonomy of communication errors," in *ISCA Tutorial and Research Workshop on Error Handling in Spoken Dialogue Systems*, 2003.

[134] M. K. Hong, A. Fourney, D. DeBellis, and S. Amershi, "Planning for natural language failures with the ai playbook," in *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 2021, pp. 1–11.

[135] S. Honig and T. Oron-Gilad, "Understanding and resolving failures in human-robot interaction: Literature review and model development," *Frontiers in psychology*, vol. 9, p. 861, 2018.

[136] M. Bajones, A. Weiss, and M. Vincze, "Help, anyone? a user study for modeling robotic behavior to mitigate malfunctions with the help of the user," *arXiv preprint arXiv:1606.02547*, 2016.

[137] N. Mirnig, G. Stollnberger, M. Miksch, S. Stadler, M. Giuliani, and M. Tscheligi, "To err is robot: How humans assess and act toward an erroneous social robot," *Frontiers in Robotics and AI*, p. 21, 2017.

[138] A. Rapp, L. Curti, and A. Boldi, "The human side of human-chatbot interaction: A systematic literature review of ten years of research on text-based chatbots," *International Journal of Human-Computer Studies*, vol. 151, p. 102 630, 2021.

[139] C.-H. Li, S.-F. Yeh, T.-J. Chang, M.-H. Tsai, K. Chen, and Y.-J. Chang, "A Conversation Analysis of Non-Progress and Coping Strategies with a Banking Task-Oriented Chatbot," in *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, New York, NY, USA: ACM, Apr. 2020, pp. 1–12, ISBN: 9781450367080.

[140] C. Myers, A. Furqan, J. Nebolsky, K. Caro, and J. Zhu, "Patterns for how users overcome obstacles in voice user interfaces," in *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, 2018, pp. 1–7.

[141] E. Luger and A. Sellen, “" like having a really bad pa" the gulf between user expectation and experience of conversational agents,” in *Proceedings of the 2016 CHI conference on human factors in computing systems*, 2016, pp. 5286–5297.

[142] E. Beneteau, O. K. Richards, M. Zhang, J. A. Kientz, J. Yip, and A. Hiniker, “Communication breakdowns between families and alexa,” in *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 2019, pp. 1–13.

[143] S. Feng, “How to convey resilience: Towards a taxonomy for conversational agent breakdown recovery strategies,” 2023.

[144] E. Alghamdi, M. Halvey, and E. Nicol, “System and user strategies to repair conversational breakdowns of spoken dialogue systems: A scoping review,” in *Proceedings of the 6th ACM Conference on Conversational User Interfaces*, 2024, pp. 1–13.

[145] D. Benner, E. Elshan, S. Schöbel, and A. Janson, “What do you mean? a review on recovery strategies to overcome conversational breakdowns of conversational agents.,” in *ICIS*, 2021.

[146] A. Mahmood, J. W. Fung, I. Won, and C.-M. Huang, “Owning mistakes sincerely: Strategies for mitigating ai errors,” in *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, 2022, pp. 1–11.

[147] M. Axelsson, M. Spitale, and H. Gunes, “" oh, sorry, i think i interrupted you": Designing repair strategies for robotic longitudinal well-being coaching,” in *Proceedings of the 2024 ACM/IEEE International Conference on Human-Robot Interaction*, 2024, pp. 13–22.

[148] J. Xu and A. Howard, “Evaluating the impact of emotional apology on human-robot trust,” in *2022 31st IEEE international conference on robot and human interactive communication (ro-man)*, IEEE, 2022, pp. 1655–1661.

[149] S. Yuan, B. Brüggemeier, S. Hillmann, and T. Michael, “User preference and categories for error responses in conversational user interfaces,” in *Proceedings of the 2nd Conference on Conversational User Interfaces*, 2020, pp. 1–8.

[150] E. S. Kox, J. H. Kerstholt, T. F. Hueting, and P. W. de Vries, “Trust repair in human-agent teams: The effectiveness of explanations and expressing regret,” *Autonomous agents and multi-agent systems*, vol. 35, no. 2, p. 30, 2021.

[151] X. Wang and M. Yin, “Are explanations helpful? a comparative study of the effects of explanations in ai-assisted decision-making,” in *26th international conference on intelligent user interfaces*, 2021, pp. 318–328.

[152] V. Chen, Q. V. Liao, J. Wortman Vaughan, and G. Bansal, "Understanding the role of human intuition on reliance in human-ai decision-making with explanations," *Proceedings of the ACM on Human-computer Interaction*, vol. 7, no. CSCW2, pp. 1–32, 2023.

[153] T. Miller, "Explanation in artificial intelligence: Insights from the social sciences," *Artificial intelligence*, vol. 267, pp. 1–38, 2019.

[154] Q. V. Liao and K. R. Varshney, "Human-centered explainable ai (xai): From algorithms to user experiences," *arXiv preprint arXiv:2110.10790*, 2021.

[155] U. Ehsan and M. O. Riedl, "Human-centered explainable ai: Towards a reflective sociotechnical approach," in *HCI International 2020-Late Breaking Papers: Multimodality and Intelligence: 22nd HCI International Conference, HCII 2020, Copenhagen, Denmark, July 19–24, 2020, Proceedings 22*, Springer, 2020, pp. 449–466.

[156] J. T. Hancock, M. Naaman, and K. Levy, "Ai-mediated communication: Definition, research agenda, and ethical considerations," *Journal of Computer-Mediated Communication*, vol. 25, no. 1, pp. 89–100, 2020.

[157] J. Hohenstein and M. Jung, "Ai as a moral crumple zone: The effects of ai-mediated communication on attribution and trust," *Computers in Human Behavior*, vol. 106, p. 106 190, 2020.

[158] H. Mieczkowski, J. T. Hancock, M. Naaman, M. Jung, and J. Hohenstein, "Ai-mediated communication: Language use and interpersonal effects in a referential communication task," *Proceedings of the ACM on Human-Computer Interaction*, vol. 5, no. CSCW1, pp. 1–14, 2021.

[159] M. Jakesch, M. French, X. Ma, J. T. Hancock, and M. Naaman, "Ai-mediated communication: How the perception that profile text was written by ai affects trustworthiness," in *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 2019, pp. 1–13.

[160] W. Rammert, *Where the action is. Distributed agency between humans, machines, and programs*. transcript, 2008.

[161] J. M. Mayer, Q. Jones, and S. R. Hiltz, "Identifying opportunities for valuable encounters: Toward context-aware social matching systems," *ACM Transactions on Information Systems (TOIS)*, vol. 34, no. 1, pp. 1–32, 2015.

[162] D. Zytko, V. Regalado, S. A. Grandhi, and Q. Jones, "Supporting online dating decisions with a prompted discussion interface," in *Companion of the 2018 ACM Conference on Computer Supported Cooperative Work and Social Computing*, 2018, pp. 353–356.

[163] M. Motoyama and G. Varghese, "I seek you: Searching and matching individuals in social networks," in *Proceedings of the eleventh international workshop on Web information and data management*, 2009, pp. 67–75.

[164] J. Chen, W. Geyer, C. Dugan, M. Muller, and I. Guy, "Make new friends, but keep the old: Recommending people on social networking sites," in *Proceedings of the SIGCHI conference on human factors in computing systems*, 2009, pp. 201–210.

[165] T. Olsson, J. Huhtamäki, and H. Kärkkäinen, "Directions for professional social matching systems," *Communications of the ACM*, vol. 63, no. 2, pp. 60–69, 2020.

[166] D. Zytko and L. DeVreugd, "Designing a social matching system to connect academic researchers with local community collaborators," *Proceedings of the ACM on Human-Computer Interaction*, vol. 3, no. GROUP, pp. 1–15, 2019.

[167] R. D. Putnam *et al.*, *Bowling alone: The collapse and revival of American community*. Simon and schuster, 2000.

[168] J. M. Mayer, S. R. Hiltz, and Q. Jones, "Making social matching context-aware: Design concepts and open challenges," in *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, 2015, pp. 545–554.

[169] J. M. Mayer, S. R. Hiltz, L. Barkhuus, K. Väänänen, and Q. Jones, "Supporting opportunities for context-aware social matching: An experience sampling study," in *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, 2016, pp. 2430–2441.

[170] E. Olshannikova, T. Olsson, J. Huhtamäki, S. Paasovaara, and H. Kärkkäinen, "From chance to serendipity: Knowledge workers' experiences of serendipitous social encounters," *Advances in Human-Computer Interaction*, vol. 2020, 2020.

[171] N. Tintarev and J. Masthoff, "Evaluating the effectiveness of explanations for recommender systems," *User Modeling and User-Adapted Interaction*, vol. 22, no. 4-5, pp. 399–439, 2012.

[172] C.-H. Tsai and P. Brusilovsky, "Providing control and transparency in a social recommender system for academic conferences," in *Proceedings of the 25th Conference on User Modeling, Adaptation and Personalization*, 2017, pp. 313–317.

[173] M. S. Ackerman, "The intellectual challenge of cscw: The gap between social requirements and technical feasibility," *Human–Computer Interaction*, vol. 15, no. 2-3, pp. 179–203, 2000.

[174] T. Erickson and W. A. Kellogg, "Social translucence: An approach to designing systems that support social processes," *ACM transactions on computer-human interaction (TOCHI)*, vol. 7, no. 1, pp. 59–83, 2000.

[175] C.-F. Chung, N. Gorm, I. A. Shklovski, and S. Munson, "Finding the right fit: Understanding health tracking in workplace wellness programs," in *Proceedings of the 2017 CHI conference on human factors in computing systems*, 2017, pp. 4875–4886.

[176] P. Duysburgh, S. A. Elprama, and A. Jacobs, "Exploring the social-technological gap in telesurgery: Collaboration within distributed or teams," in *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing*, 2014, pp. 1537–1548.

[177] A. Gheitasy, J. Abdelnour-Nocera, and B. Nardi, "Socio-technical gaps in online collaborative consumption (occ) an example of the etsy community," in *Proceedings of the 33rd Annual International Conference on the Design of Communication*, 2015, pp. 1–9.

[178] W. A. Kellogg and T. Erickson, "Social translucence, collective awareness, and the emergence of place," *Proceedings of CSCW2002*, pp. 1–6, 2002.

[179] D. W. McDonald, S. Gokhman, and M. Zachry, "Building for social translucence: A domain analysis and prototype system," in *Proceedings of the ACM 2012 conference on computer supported cooperative work*, 2012, pp. 637–646.

[180] A. M. Szostek, E. Karapanos, B. Eggen, and M. Holenderski, "Understanding the implications of social translucence for systems supporting communication at work," in *Proceedings of the 2008 ACM conference on Computer supported cooperative work*, 2008, pp. 649–658.

[181] P. Bjørn and O. Ngwenyama, "Virtual team collaboration: Building shared meaning, resolving breakdowns and creating translucence," *Information systems journal*, vol. 19, no. 3, pp. 227–253, 2009.

[182] J. Byun, J. Park, and A. Oh, "Cocode: Co-learner screen sharing for social translucence in online programming courses," in *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*, 2020, pp. 1–4.

[183] M. Ito, *Hanging out, messing around, and geeking out: Kids living and learning with new media*. The MIT Press, 2013.

[184] C. Hostetter and M. Busch, "Measuring up online: The relationship between social presence and student learning satisfaction," *Journal of the Scholarship of Teaching and Learning*, pp. 1–12, 2006.

[185] A. P. Rovai, "Building classroom community at a distance: A case study," *Educational technology research and development*, vol. 49, no. 4, p. 33, 2001.

[186] A. P. Rovai, "Development of an instrument to measure classroom community," *The Internet and higher education*, vol. 5, no. 3, pp. 197–211, 2002.

[187] T. I. Aldosemani, C. E. Shepherd, I. Gashim, and T. Dousay, "Developing third places to foster sense of community in online instruction," *British Journal of Educational Technology*, vol. 47, no. 6, pp. 1020–1031, 2016.

[188] S. R. Aragon, "Creating social presence in online environments," *New directions for adult and continuing education*, vol. 2003, no. 100, pp. 57–68, 2003.

[189] C. M. Johnson, "A survey of current research on online communities of practice," *The internet and higher education*, vol. 4, no. 1, pp. 45–60, 2001.

[190] C. Irwin and Z. Berge, "Socialization in the online classroom.," *E-Journal of Instructional Science and Technology*, vol. 9, no. 1, n1, 2006.

[191] K. Kreijns, P. A. Kirschner, and W. Jochems, "Identifying the pitfalls for social interaction in computer-supported collaborative learning environments: A review of the research," *Computers in human behavior*, vol. 19, no. 3, pp. 335–353, 2003.

[192] D. Yoon *et al.*, "Richreview++ deployment of a collaborative multi-modal annotation system for instructor feedback and peer discussion," in *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing*, 2016, pp. 195–205.

[193] H. J. Teo and A. Johri, "Fast, functional, and fitting: Expert response dynamics and response quality in an online newcomer help forum," in *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing*, 2014, pp. 332–341.

[194] A. Cross, M. Bayyapunedi, D. Ravindran, E. Cutrell, and W. Thies, "Vidwiki: Enabling the crowd to improve the legibility of online educational videos," in *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing*, 2014, pp. 1167–1175.

[195] R. Ahuja, D. Khan, D. Symonette, M. desJardins, S. Stacey, and D. Engel, "A digital dashboard for supporting online student teamwork," in *Conference Companion Publication of the 2019 on Computer Supported Cooperative Work and Social Computing*, 2019, pp. 132–136.

[196] N. Sun, M. B. Rosson, and J. M. Carroll, "Where is community among online learners? identity, efficacy and personal ties," in *Proceedings of the 2018 chi conference on human factors in computing systems*, 2018, pp. 1–13.

[197] G. Siemens and R. S. d. Baker, "Learning analytics and educational data mining: Towards communication and collaboration," in *Proceedings of the 2nd international conference on learning analytics and knowledge*, 2012, pp. 252–254.

[198] X. Du, J. Yang, B. E. Shelton, J.-L. Hung, and M. Zhang, "A systematic meta-review and analysis of learning analytics research," *Behaviour & Information Technology*, vol. 40, no. 1, pp. 49–62, 2021.

[199] M. C. Goulden *et al.*, "Ccvis: Visual analytics of student online learning behaviors using course clickstream data," *Electronic Imaging*, vol. 2019, no. 1, pp. 681–1, 2019.

[200] Q. Li, R. Baker, and M. Warschauer, "Using clickstream data to measure, understand, and support self-regulated learning in online courses," *The Internet and Higher Education*, vol. 45, p. 100 727, 2020.

[201] Z. Zhang, Z. Li, H. Liu, T. Cao, and S. Liu, "Data-driven online learning engagement detection via facial expression and mouse behavior recognition technology," *Journal of Educational Computing Research*, vol. 58, no. 1, pp. 63–86, 2020.

[202] L. Wang and Y. Yuan, "A prediction strategy for academic records based on classification algorithm in online learning environment," in *2019 IEEE 19th International Conference on Advanced Learning Technologies (ICALT)*, IEEE, vol. 2161, 2019, pp. 1–5.

[203] S. L. Hoskins and J. C. Van Hooff, "Motivation and ability: Which students use online learning and what influence does it have on their achievement?" *British journal of educational technology*, vol. 36, no. 2, pp. 177–192, 2005.

[204] M. Herbert, "Staying the course: A study in online student satisfaction and retention," *Online Journal of Distance Learning Administration*, vol. 9, no. 4, pp. 300–317, 2006.

[205] M. Layne, W. E. Boston, and P. Ice, "A longitudinal study of online learners: Shoppers, swirlers, stoppers, and succeeders as a function of demographic characteristics," *Online Journal of Distance Learning Administration*, vol. 16, no. 2, pp. 1–12, 2013.

[206] I. Camacho and A. Goel, "Longitudinal trends in sentiment polarity and readability of an online masters of computer science course," in *Proceedings of the Fifth Annual ACM Conference on Learning at Scale*, 2018, pp. 1–4.

[207] J. W. You, "Identifying significant indicators using lms data to predict course achievement in online learning," *The Internet and Higher Education*, vol. 29, pp. 23–30, 2016.

[208] I. Irish *et al.*, "Parqr: Automatic post suggestion in the piazza online forum to support degree seeking online masters students," in *Proceedings of the Seventh ACM Conference on Learning@ Scale*, 2020, pp. 125–134.

[209] K. M. Al-Aubidy, "Applying fuzzy logic for learner modeling and decision support in online learning systems.," *Journal of Educational Technology*, vol. 2, no. 3, pp. 76–85, 2005.

[210] S. Slade and P. Prinsloo, "Learning analytics: Ethical issues and dilemmas," *American Behavioral Scientist*, vol. 57, no. 10, pp. 1510–1529, 2013.

[211] K. Sun, A. H. Mhaidli, S. Watel, C. A. Brooks, and F. Schaub, "It's my data! tensions among stakeholders of a learning analytics dashboard," in *Proceedings of the 2019 chi conference on human factors in computing systems*, 2019, pp. 1–14.

[212] S. Slade, P. Prinsloo, and M. Khalil, "Learning analytics at the intersections of student trust, disclosure and benefit," in *Proceedings of the 9th International Conference on learning analytics & knowledge*, 2019, pp. 235–244.

[213] J. Whitmer, K. Fernandes, and W. R. Allen, "Analytics in progress: Technology use, student characteristics, and student achievement," *EDUCAUSE Review Online*, vol. 7, 2012.

[214] L. Taber and S. Whittaker, """ on finsta, i can say'hail satan'": Being authentic but disagreeable on instagram," in *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 2020, pp. 1–14.

[215] X. Huang, J. Vitak, and Y. Tausczik, "" you don't have to know my past": How wechat moments users manage their evolving self-presentation," in *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 2020, pp. 1–13.

[216] S. Chancellor, M. L. Birnbaum, E. D. Caine, V. M. Silenzio, and M. De Choudhury, "A taxonomy of ethical tensions in inferring mental health states from social media," in *Proceedings of the conference on fairness, accountability, and transparency*, 2019, pp. 79–88.

[217] H. Park and J. Lee, "Designing a conversational agent for sexual assault survivors: Defining burden of self-disclosure and envisioning survivor-centered solutions," in *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 2021, pp. 1–17.

[218] Z. Xiao *et al.*, "Tell me about yourself: Using an ai-powered chatbot to conduct conversational surveys with open-ended questions," *ACM Transactions on Computer-Human Interaction (TOCHI)*, vol. 27, no. 3, pp. 1–37, 2020.

[219] W. Xu, "Toward human-centered ai: A perspective from human-computer interaction," *interactions*, vol. 26, no. 4, pp. 42–46, 2019.

[220] J. B. Arbaugh *et al.*, "Developing a community of inquiry instrument: Testing a measure of the community of inquiry framework using a multi-institutional sample," *The internet and higher education*, vol. 11, no. 3-4, pp. 133–136, 2008.

[221] J. Lave, E. Wenger, *et al.*, *Situated learning: Legitimate peripheral participation*. Cambridge university press, 1991.

[222] O. E. Nordberg *et al.*, "Designing chatbots for guiding online peer support conversations for adults with adhd," in *International Workshop on Chatbot Research and Design*, Springer, 2019, pp. 113–126.

[223] J. Narain, T. Quach, M. Davey, H. W. Park, C. Breazeal, and R. Picard, "Promoting wellbeing with sunny, a chatbot that facilitates positive messages within social groups," in *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*, 2020, pp. 1–8.

[224] L. A. Palinkas, S. M. Horwitz, C. A. Green, J. P. Wisdom, N. Duan, and K. Hoagwood, "Purposeful sampling for qualitative data collection and analysis in mixed method implementation research," *Administration and policy in mental health and mental health services research*, vol. 42, no. 5, pp. 533–544, 2015.

[225] K. Charmaz, *Constructing grounded theory*. sage, 2014.

[226] J. Larreamendy-Joerns and G. Leinhardt, "Going the distance with online education," *Review of educational research*, vol. 76, no. 4, pp. 567–605, 2006.

[227] C. Fiesler, C. Lampe, and A. S. Bruckman, "Reality and perception of copyright terms of service for online content creation," in *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing*, 2016, pp. 1450–1461.

[228] J. Grudin, "Groupware and social dynamics: Eight challenges for developers," *Communications of the ACM*, vol. 37, no. 1, pp. 92–105, 1994.

[229] S. Milano, M. Taddeo, and L. Floridi, "Recommender systems and their ethical challenges," *AI & SOCIETY*, vol. 35, no. 4, pp. 957–967, 2020.

[230] D. W. Yoo and M. De Choudhury, "Designing dashboard for campus stakeholders to support college student mental health," in *Proceedings of the 13th EAI International Conference on Pervasive Computing Technologies for Healthcare*, 2019, pp. 61–70.

[231] K. Saha *et al.*, "Person-centered predictions of psychological constructs with social media contextualized by multimodal sensing," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 5, no. 1, pp. 1–32, 2021.

[232] A. Acquisti, L. Brandimarte, and G. Loewenstein, "Privacy and human behavior in the age of information," *Science*, vol. 347, no. 6221, pp. 509–514, 2015.

[233] U. Ehsan, Q. V. Liao, M. Muller, M. O. Riedl, and J. D. Weisz, "Expanding explainability: Towards social transparency in ai systems," in *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 2021, pp. 1–19.

[234] C. J. Cai, S. Winter, D. Steiner, L. Wilcox, and M. Terry, "" hello ai": Uncovering the onboarding needs of medical practitioners for human-ai collaborative decision-making," *Proceedings of the ACM on Human-computer Interaction*, vol. 3, no. CSCW, pp. 1–24, 2019.

[235] Q. V. Liao, D. Gruen, and S. Miller, "Questioning the ai: Informing design practices for explainable ai user experiences," in *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 2020, pp. 1–15.

[236] S. Jhaver, I. Birman, E. Gilbert, and A. Bruckman, "Human-machine collaboration for content regulation: The case of reddit automoderator," *ACM Transactions on Computer-Human Interaction (TOCHI)*, vol. 26, no. 5, pp. 1–35, 2019.

[237] N. Eagle and A. Pentland, "Social serendipity: Mobilizing social software," *IEEE Pervasive computing*, vol. 4, no. 2, pp. 28–34, 2005.

[238] Y.-S. Chiu, K.-H. Lin, and J.-S. Chen, "A social network-based serendipity recommender system," in *2011 International Symposium on Intelligent Signal Processing and Communications Systems (ISPACS)*, IEEE, 2011, pp. 1–5.

[239] C. Brown, C. Efstratiou, I. Leontiadis, D. Quercia, and C. Mascolo, "Tracking serendipitous interactions: How individual cultures shape the office," in *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing*, 2014, pp. 1072–1081.

[240] C. Holotescu, "Moocbuddy: A chatbot for personalized learning with moocs.," in *RoCHI*, 2016, pp. 91–94.

[241] W. Huang, K. F. Hew, and D. E. Gonda, "Designing and evaluating three chatbot-enhanced activities for a flipped graduate course," *International Journal of Mechanical Engineering and Robotics Research*, vol. 8, no. 5, p. 6, 2019.

[242] J. Lu, D. Wu, M. Mao, W. Wang, and G. Zhang, "Recommender system application developments: A survey," *Decision Support Systems*, vol. 74, pp. 12–32, 2015.

[243] M. Lee, L. Frank, F. Beute, Y. De Kort, and W. IJsselsteijn, "Bots mind the social-technical gap," in *Proceedings of 15th European conference on computer-supported cooperative work-exploratory papers*, European Society for Socially Embedded Technologies (EUSSET), 2017.

[244] L. Ring, B. Barry, K. Totzke, and T. Bickmore, "Addressing loneliness and isolation in older adults: Proactive affective agents provide better support," in *2013 Humaine Association conference on affective computing and intelligent interaction*, IEEE, 2013, pp. 61–66.

[245] L. L. Kramer, M. Blok, L. Van Velsen, B. C. Mulder, and E. De Vet, "Supporting eating behaviour of community-dwelling older adults: Co-design of an embodied conversational agent," *Design for Health*, pp. 1–20, 2021.

[246] K. K. Fitzpatrick, A. Darcy, and M. Vierhile, "Delivering cognitive behavior therapy to young adults with symptoms of depression and anxiety using a fully automated conversational agent (woebot): A randomized controlled trial," *JMIR mental health*, vol. 4, no. 2, e7785, 2017.

[247] A. P. Chaves and M. A. Gerosa, "How should my chatbot interact? a survey on social characteristics in human–chatbot interaction design," *International Journal of Human–Computer Interaction*, vol. 37, no. 8, pp. 729–758, 2021.

[248] M. De Gennaro, E. G. Krumhuber, and G. Lucas, "Effectiveness of an empathic chatbot in combating adverse effects of social exclusion on mood," *Frontiers in psychology*, vol. 10, p. 3061, 2020.

[249] C. Falala-Séchet, L. Antoine, I. Thiriez, and C. Bungener, "Owlie: A chatbot that provides emotional support for coping with psychological difficulties," in *Proceedings of the 19th ACM International Conference on Intelligent Virtual Agents*, 2019, pp. 236–237.

[250] E. Svikhnushina and P. Pu, "Social and emotional etiquette of chatbots: A qualitative approach to understanding user needs and expectations," *arXiv preprint arXiv:2006.13883*, 2020.

[251] K.-J. Oh, D. Lee, B. Ko, and H.-J. Choi, "A chatbot for psychiatric counseling in mental healthcare service based on emotional dialogue analysis and sentence gen-

eration," in *2017 18th IEEE International Conference on Mobile Data Management (MDM)*, IEEE, 2017, pp. 371–375.

[252] D. Lee, K.-J. Oh, and H.-J. Choi, "The chatbot feels you-a counseling service using emotional response generation," in *2017 IEEE international conference on big data and smart computing (BigComp)*, IEEE, 2017, pp. 437–440.

[253] E. B.-N. Sanders and P. J. Stappers, "Co-creation and the new landscapes of design," *Co-design*, vol. 4, no. 1, pp. 5–18, 2008.

[254] T. Robertson and J. Simonsen, "Participatory design: An introduction," in *Routledge international handbook of participatory design*, Routledge, 2012, pp. 21–38.

[255] Z. Chen, Y. Lu, M. P. Nieminen, and A. Lucero, "Creating a chatbot for and with migrants: Chatbot personality drives co-design activities," in *Proceedings of the 2020 ACM Designing Interactive Systems Conference*, 2020, pp. 219–230.

[256] L. S. G. Piccolo, P. Troullinou, and H. Alani, "Chatbots to support children in coping with online threats: Socio-technical requirements," in *Designing Interactive Systems Conference 2021*, 2021, pp. 1504–1517.

[257] R. Garg and S. Sengupta, "Conversational technologies for in-home learning: Using co-design to understand children's and parents' perspectives," in *Proceedings of the 2020 CHI conference on human factors in computing systems*, 2020, pp. 1–13.

[258] E. A. Björling and E. Rose, "Participatory research principles in human-centered design: Engaging teens in the co-design of a social robot," *Multimodal Technologies and Interaction*, vol. 3, no. 1, p. 8, 2019.

[259] M. Luria, O. Sheriff, M. Boo, J. Forlizzi, and A. Zoran, "Destruction, catharsis, and emotional release in human-robot interaction," *ACM Transactions on Human-Robot Interaction (THRI)*, vol. 9, no. 4, pp. 1–19, 2020.

[260] H. Beyer and K. Holtzblatt, "Contextual design," *interactions*, vol. 6, no. 1, pp. 32–42, 1999.

[261] J. Woodward, Z. McFadden, N. Shiver, A. Ben-hayon, J. C. Yip, and L. Anthony, "Using co-design to examine how children conceptualize intelligent interfaces," in *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, 2018, pp. 1–14.

[262] Y.-C. Lee, N. Yamashita, Y. Huang, and W. Fu, ""' i hear you, i feel you": Encouraging deep self-disclosure through a chatbot," in *Proceedings of the 2020 CHI conference on human factors in computing systems*, 2020, pp. 1–12.

[263] S. Kim, J. Eun, C. Oh, B. Suh, and J. Lee, "Bot in the bunch: Facilitating group chat discussion by improving efficiency and participation with a chatbot," in *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 2020, pp. 1–13.

[264] T. L. Smestad and F. Volden, "Chatbot personalities matters," in *International Conference on Internet Science*, Springer, 2018, pp. 170–181.

[265] D. Gray, *Challenge cards*, https://gamestorming.com/challenge-cards/, Accessed: 2021-08-19, 2011.

[266] L. Ciechanowski, A. Przegalinska, M. Magnuski, and P. Gloor, "In the shades of the uncanny valley: An experimental study of human–chatbot interaction," *Future Generation Computer Systems*, vol. 92, pp. 539–548, 2019.

[267] A. Xu, Z. Liu, Y. Guo, V. Sinha, and R. Akkiraju, "A new chatbot for customer service on social media," in *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, 2017, pp. 3506–3510.

[268] S. Ruan *et al.*, "Quizbot: A dialogue-based adaptive learning system for factual knowledge," in *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 2019, pp. 1–13.

[269] L. Clark *et al.*, "What makes a good conversation? Challenges in designing truly conversational agents," *Conference on Human Factors in Computing Systems - Proceedings*, pp. 1–12, 2019.

[270] J. Seering, M. Luria, G. Kaufman, and J. Hammer, "Beyond dyadic interactions: Considering chatbots as community members," in *Conference on Human Factors in Computing Systems - Proceedings*, 2019, ISBN: 9781450359702.

[271] S. Amershi *et al.*, "Guidelines for human-ai interaction," in *Proceedings of the 2019 chi conference on human factors in computing systems*, 2019, pp. 1–13.

[272] E. Luger and A. Sellen, ""Like having a really bad PA": the gulf between user expectation and experience of conversational agents," *CHI '16 Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, pp. 5286–5297, 2016.

[273] J. Zamora, "I'm sorry, dave, i'm afraid i can't do that: Chatbot perception and expectations," in *Proceedings of the 5th International Conference on Human Agent Interaction*, 2017, pp. 253–260.

[274] O. E. Nordberg *et al.*, "Designing Chatbots for Guiding Online Peer Support Conversations for Adults with ADHD," *Lecture Notes in Computer Science (including*

subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), vol. 11970 LNCS, no. November, pp. 113–126, 2020.

[275] R. Winkler, S. Hobert, A. Salovaara, M. Söllner, and J. M. Leimeister, "Sara, the Lecturer: Improving Learning in Online Education with a Scaffolding-Based Conversational Agent," in *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, New York, NY, USA: ACM, Apr. 2020, pp. 1–14, ISBN: 9781450367080.

[276] J. Seering, J. P. Flores, S. Savage, and J. Hammer, "The social roles of bots: Situating bots in discussions in online communities," *Proceedings of the ACM on Human-Computer Interaction*, vol. 2, no. CSCW, 2018.

[277] S. Kim, J. Eun, C. Oh, B. Suh, and J. Lee, "Bot in the Bunch: Facilitating Group Chat Discussion by Improving Efficiency and Participation with a Chatbot," in *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, New York, NY, USA: ACM, Apr. 2020, pp. 1–13, ISBN: 9781450367080.

[278] J. T. Hancock, K. Gee, K. Ciaccio, and J. M.-H. Lin, "I'm sad you're sad: Emotional contagion in cmc," in *Proceedings of the 2008 ACM conference on Computer supported cooperative work*, 2008, pp. 295–298.

[279] V. A. Aleven and K. R. Koedinger, "An effective metacognitive strategy: Learning by doing and explaining with a computer-based cognitive tutor," *Cognitive science*, vol. 26, no. 2, pp. 147–179, 2002.

[280] C. Dede, J. Richards, and B. Saxberg, *Learning Engineering for Online Education: Theoretical Contexts and Design-based Examples*. Routledge, 2018.

[281] A. Goel, "Ai-powered learning: Making education accessible, affordable, and achievable," *arXiv preprint arXiv:2006.01908*, 2020.

[282] C. Bartneck, D. Kulić, E. Croft, and S. Zoghbi, "Measurement Instruments for the Anthropomorphism, Animacy, Likeability, Perceived Intelligence, and Perceived Safety of Robots," *International Journal of Social Robotics*, vol. 1, no. 1, pp. 71–81, Jan. 2009.

[283] Y. Jeong, Y. Kang, and J. Lee, "Exploring effects of conversational fillers on user perception of conversational agents," *Conference on Human Factors in Computing Systems - Proceedings*, pp. 1–6, 2019.

[284] P. E. McKight and J. Najab, "Kruskal-wallis test," *The corsini encyclopedia of psychology*, pp. 1–1, 2010.

[285] E. Go and S. S. Sundar, "Humanizing chatbots: The effects of visual, identity and conversational cues on humanness perceptions," *Computers in Human Behavior*, vol. 97, pp. 304–316, 2019.

[286] J. Cassell and T. Bickmore, "External manifestations of trustworthiness in the interface," *Communications of the ACM*, vol. 43, no. 12, pp. 50–56, 2000.

[287] A. Kuzminykh, J. Sun, N. Govindaraju, J. Avery, and E. Lank, "Genie in the Bottle: Anthropomorphized Perceptions of Conversational Agents," in *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, New York, NY, USA: ACM, Apr. 2020, pp. 1–13, ISBN: 9781450367080.

[288] X. Yang, M. Aurisicchio, and W. Baxter, "Understanding Affective Experiences with Conversational Agents," in *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems - CHI '19*, New York, New York, USA: ACM Press, 2019, pp. 1–12, ISBN: 9781450359702.

[289] M. Mori, K. F. MacDorman, and N. Kageki, "The uncanny valley [from the field]," *IEEE Robotics & Automation Magazine*, vol. 19, no. 2, pp. 98–100, 2012.

[290] E. J. De Visser *et al.*, "Almost human: Anthropomorphism increases trust resilience in cognitive agents.," *Journal of Experimental Psychology: Applied*, vol. 22, no. 3, p. 331, 2016.

[291] M. D. Pickard, J. K. Burgoon, and D. C. Derrick, "Toward an Objective Linguistic-Based Measure of Perceived Embodied Conversational Agent Power and Likeability," *International Journal of Human-Computer Interaction*, vol. 30, no. 6, pp. 495–516, 2014.

[292] S. Reysen, "Construction of a new scale: The reysen likability scale," *Social Behavior and Personality: an international journal*, vol. 33, no. 2, pp. 201–208, 2005.

[293] T. L. Robbins and A. S. DeNisi, "A closer look at interpersonal affect as a distinct influence on cognitive processing in performance evaluations.," *Journal of Applied Psychology*, vol. 79, no. 3, p. 341, 1994.

[294] R. M. Dawes and B. Corrigan, "Linear models in decision making.," *Psychological bulletin*, vol. 81, no. 2, p. 95, 1974.

[295] K. Saha and A. Sharma, "Causal Factors of Effective Psychosocial Outcomes in Online Mental Health Communities," in *ICWSM*, 2020.

[296] S. K. Ernala, A. F. Rizvi, M. L. Birnbaum, J. M. Kane, and M. De Choudhury, "Linguistic markers indicating therapeutic outcomes of social media disclosures of

schizophrenia," *Proceedings of the ACM on Human-Computer Interaction*, vol. 1, no. CSCW, pp. 1–27, 2017.

[297] N. A. Murphy, "Appearing smart: The impression management of intelligence, person perception accuracy, and behavior in social interaction," *Personality and Social Psychology Bulletin*, vol. 33, no. 3, pp. 325–339, 2007.

[298] J. Hill, W. Randolph Ford, and I. G. Farreras, "Real conversations with artificial intelligence: A comparison between human-human online conversations and human-chatbot conversations," *Computers in Human Behavior*, vol. 49, pp. 245–250, 2015.

[299] C. L. Lortie and M. J. Guitton, "Judgment of the humanness of an interlocutor is in the eye of the beholder," *PLoS One*, vol. 6, no. 9, e25085, 2011.

[300] H. Fang *et al.*, "Sounding board: A user-centric and content-driven social chatbot," *arXiv preprint arXiv:1804.10202*, 2018.

[301] D. R. McCallum and J. L. Peterson, "Computer-based readability indexes," in *Proceedings of the ACM'82 Conference*, 1982, pp. 44–48.

[302] K. Saha, I. Weber, and M. De Choudhury, "A Social Media Based Examination of the Effects of Counseling Recommendations After Student Deaths on College Campuses," in *ICWSM*, 2018.

[303] E. Pitler and A. Nenkova, "Revisiting readability: A unified framework for predicting text quality," in *Proceedings of the conference on empirical methods in natural language processing*, Association for Computational Linguistics, 2008, pp. 186–195.

[304] J. Feine, S. Morana, and U. Gnewuch, "Measuring Service Encounter Satisfaction with Customer Service Chatbots using Sentiment Analysis," *Proceedings of the 14th International Conference on Wirtschaftsinformatik*, no. December, pp. 0–11, 2019.

[305] N. Novielli, F. de Rosis, and I. Mazzotta, "User attitude towards an embodied conversational agent: Effects of the interaction mode," *Journal of Pragmatics*, vol. 42, no. 9, pp. 2385–2397, 2010.

[306] C. J. Hutto and E. E. Gilbert, "VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text. Eighth International Conference on Weblogs and Social Media (ICWSM-14).","*Proceedings of the 8th International Conference on Weblogs and Social Media, ICWSM 2014*, 2014.

[307] T. Althoff, K. Clark, and J. Leskovec, "Large-scale analysis of counseling conversations: An application of natural language processing to mental health," *TACL*, 2016.

[308] J. Pennington, R. Socher, and C. Manning, "Glove: Global vectors for word representation," in *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 2014, pp. 1532–1543.

[309] T. Mikolov, I. Sustskever, K. Chen, G. Corrado, and J. Dean, "Distributed Representations ofWords and Phrases and their Compositionality," in *Neural Information Processing Systems (NIPS)*, 2013, pp. 3111–3119.

[310] V. Das Swain, K. Saha, M. D. Reddy, H. Rajvanshy, G. D. Abowd, and M. De Choudhury, "Modeling Organizational Culture with Workplace Experiences Shared on Glassdoor," in *CHI*, 2020.

[311] C. Nass and Y. Moon, "Machines and mindlessness: Social responses to computers.," *Journal of Social Issues*, vol. 1, no. 56, pp. 81–103, 2000.

[312] P. Grimm, "Social desirability bias," *Wiley international encyclopedia of marketing*, 2010.

[313] E. Chandrasekharan, M. Samory, A. Srinivasan, and E. Gilbert, "The Bag of Communities: Identifying Abusive Behavior Online with Preexisting Internet Data," in *Proc. CHI*, 2017.

[314] D. Choi, D. Kwak, M. Cho, and S. Lee, ""Nobody Speaks that Fast!" An Empirical Study of Speech Rate in Conversational Agents for People with Vision Impairments," in *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, New York, NY, USA: ACM, Apr. 2020, pp. 1–13, ISBN: 9781450367080.

[315] J. Seering, M. Luria, C. Ye, G. Kaufman, and J. Hammer, "It Takes a Village: Integrating an Adaptive Chatbot into an Online Gaming Community," in *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, New York, NY, USA: ACM, Apr. 2020, pp. 1–13, ISBN: 9781450367080.

[316] D.-j. Kim and Y.-k. Lim, "Co-Performing Agent: Design for Building User-Agent Partnership in Learning and Adaptive Services," in *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems - CHI '19*, New York, New York, USA: ACM Press, 2019, pp. 1–14, ISBN: 9781450359702.

[317] K. Jaidka, S. C. Guntuku, A. Buffone, H. A. Schwartz, and L. H. Ungar, "Facebook vs. twitter: Cross-platform differences in self-disclosure and trait prediction,"

in *Proceedings of the Twelfth International AAAI Conference on Web and Social Media*, 2018, pp. 141–150.

[318] C. Danescu-Niculescu-Mizil, M. Gamon, and S. Dumais, "Mark my words! linguistic style accommodation in social media," in *Proceedings of the 20th international conference on World wide web*, 2011, pp. 745–754.

[319] K. Saha, M. D. Reddy, S. Mattingly, E. Moskal, A. Sirigiri, and M. De Choudhury, "Libra: On linkedin based role ambiguity and its relationship with wellbeing and job performance," *Proceedings of the ACM on Human-Computer Interaction*, vol. 3, no. CSCW, pp. 1–30, 2019.

[320] Z. Xiao, M. X. Zhou, and W.-T. Fu, "Who should be my teammates: Using a conversational agent to understand individuals and help teaming," in *Proceedings of the 24th International Conference on Intelligent User Interfaces*, 2019, pp. 437–447.

[321] J. M. Alberola, E. Del Val, V. Sanchez-Anguix, A. Palomares, and M. D. Teruel, "An artificial intelligence tool for heterogeneous team formation in the classroom," *Knowledge-Based Systems*, vol. 101, pp. 1–14, 2016.

[322] F. Jahanbakhsh, W.-T. Fu, K. Karahalios, D. Marinov, and B. Bailey, "You want me to work with who? stakeholder perceptions of automated team formation in project-based courses," in *Proceedings of the 2017 CHI conference on human factors in computing systems*, 2017, pp. 3201–3212.

[323] I. Lykourentzou, A. Antoniou, Y. Naudet, and S. P. Dow, "Personality matters: Balancing for personality types leads to better outcomes for crowd teams," in *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing*, 2016, pp. 260–273.

[324] M. Salem, G. Lakatos, F. Amirabdollahian, and K. Dautenhahn, "Would you trust a (faulty) robot? effects of error, task type and personality on human-robot cooperation and trust," in *2015 10th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, IEEE, 2015, pp. 1–8.

[325] S. Gulati, S. Sousa, and D. Lamas, "Design, development and evaluation of a human-computer trust scale," *Behaviour & Information Technology*, vol. 38, no. 10, pp. 1004–1015, 2019.

[326] V. Braun and V. Clarke, "Using thematic analysis in psychology," *Qualitative research in psychology*, vol. 3, no. 2, pp. 77–101, 2006.

[327] V. Braun and V. Clarke, "Reflecting on reflexive thematic analysis," *Qualitative research in sport, exercise and health*, vol. 11, no. 4, pp. 589–597, 2019.

[328] D. Dickson and I. Kelly, "The 'barnum effect'in personality assessment: A review of the literature," *Psychological reports*, vol. 57, no. 2, pp. 367–382, 1985.

[329] C. Chen, S. Feng, A. Sharma, and C. Tan, "Machine explanations and human understanding," in *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, 2023, pp. 1–1.

[330] J. Schoeffer, N. Kuehl, and Y. Machowski, ""there is not enough information": On the effects of explanations on perceptions of informational fairness and trustworthiness in automated decision-making," in *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, 2022, pp. 1616–1628.

[331] M. Eslami *et al.*, "First i" like" it, then i hide it: Folk theories of social feeds," in *Proceedings of the 2016 cHI conference on human factors in computing systems*, 2016, pp. 2371–2382.

[332] D. Long and B. Magerko, "What is ai literacy? competencies and design considerations," in *Proceedings of the 2020 CHI conference on human factors in computing systems*, 2020, pp. 1–16.

[333] S. Druga, R. Williams, C. Breazeal, and M. Resnick, "" hey google is it ok if i eat you?" initial explorations in child-agent interaction," in *Proceedings of the 2017 conference on interaction design and children*, 2017, pp. 595–600.

[334] M. K. Lee, S. Kiesler, J. Forlizzi, S. Srinivasa, and P. Rybski, "Gracefully mitigating breakdowns in robotic services," in *2010 5th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, IEEE, 2010, pp. 203–210.

[335] M. T. Cox, "Metacognition in computation: A selected research review," *Artificial intelligence*, vol. 169, no. 2, pp. 104–141, 2005.

[336] M. Cox and A. Raja, "Metareasoning: A manifesto," *BBN Technical*, 2007.

[337] M. B. Ganapini *et al.*, "Thinking fast and slow in ai: The role of metacognition," in *International Conference on Machine Learning, Optimization, and Data Science*, Springer, 2022, pp. 502–509.

[338] D. Kahneman, *Thinking, fast and slow*. macmillan, 2011.

[339] M. Schmill *et al.*, "The role of metacognition in robust ai systems," in *Workshop on Metareasoning at the Twenty-Third AAAI Conference on Artificial Intelligence*, 2008.

[340] A. Sloman, *Varieties of metacognition in natural and artificial systems.* 2011.

[341] A. Goel, A. G. de Silver Garza, N. Grué, J. W. Murdock, M. Recker, and T. Govindaraj, "Explanatory interface in interactive design environments," *Artificial intelligence in design'96*, pp. 387–405, 1996.

[342] J. W. Murdock and A. K. Goel, "Meta-case-based reasoning: Self-improvement through self-understanding," *Journal of Experimental & Theoretical Artificial Intelligence*, vol. 20, no. 1, pp. 1–36, 2008.

[343] A. Goel, H. Sikka, V. Nandan, J. Lee, M. Lisle, and S. Rugaber, "Explanation as question answering based on a task model of the agent's design," *arXiv preprint arXiv:2206.05030*, 2022.

[344] R. Basappa, M. Tekman, H. Lu, B. Faught, S. Kakar, and A. K. Goel, "Social ai agents too need to explain themselves," in *International Conference on Intelligent Tutoring Systems*, Springer, 2024, pp. 351–360.

[345] T. Paek and E. J. Horvitz, "Conversation as action under uncertainty," *arXiv preprint arXiv:1301.3883*, 2013.

[346] M. Desai, P. Kaniarasu, M. Medvedev, A. Steinfeld, and H. Yanco, "Impact of robot failures and feedback on real-time trust," in *2013 8th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, IEEE, 2013, pp. 251–258.

[347] A. Cuadra, S. Li, H. Lee, J. Cho, and W. Ju, "My bad! repairing intelligent voice assistant errors improves interaction," *Proceedings of the ACM on Human-Computer Interaction*, vol. 5, no. CSCW1, pp. 1–24, 2021.

[348] J. Zeng, D. Fan, X. Zhou, and J. Tang, "Chatbot with resilience: The impact of repair strategies on customer satisfaction in conversational breakdowns," in *Wuhan International Conference on E-business*, Springer, 2024, pp. 306–317.

[349] M. T. Dzindolet, S. A. Peterson, R. A. Pomranky, L. G. Pierce, and H. P. Beck, "The role of trust in automation reliance," *International journal of human-computer studies*, vol. 58, no. 6, pp. 697–718, 2003.

[350] R. Zhang *et al.*, "I know this looks bad, but i can explain: Understanding when ai should explain actions in human-ai teams," *ACM Transactions on Interactive Intelligent Systems*, vol. 14, no. 1, pp. 1–23, 2024.

[351] C. P. Lee, P. Praveena, and B. Mutlu, "Rex: Designing user-centered repair and explanations to address robot failures," in *Proceedings of the 2024 ACM Designing Interactive Systems Conference*, 2024, pp. 2911–2925.

[352] K. Hald, K. Weitz, E. André, and M. Rehm, ""an error occurred!"-trust repair with virtual robot using levels of mistake explanation," in *Proceedings of the 9th International Conference on Human-Agent Interaction*, 2021, pp. 218–226.

[353] B. L. Pompe, E. Velner, and K. P. Truong, "The robot that showed remorse: Repairing trust with a genuine apology," in *2022 31st IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*, IEEE, 2022, pp. 260–265.

[354] T. Kim and H. Song, "How should intelligent agents apologize to restore trust? interaction effects between anthropomorphism and apology attribution on trust repair," *Telematics and Informatics*, vol. 61, p. 101 595, 2021.

[355] X. Zhang, S. K. Lee, W. Kim, and S. Hahn, ""sorry, it was my fault": Repairing trust in human-robot interactions," *International Journal of Human-Computer Studies*, vol. 175, p. 103 031, 2023.

[356] T. Jensen, Y. Albayram, M. M. H. Khan, M. A. A. Fahim, R. Buck, and E. Coman, "The apple does fall far from the tree: User separation of a system from its developers in human-automation trust repair," in *Proceedings of the 2019 on Designing Interactive Systems Conference*, 2019, pp. 1071–1082.

[357] H. Aguinis and K. J. Bradley, "Best practice recommendations for designing and implementing experimental vignette methodology studies," *Organizational research methods*, vol. 17, no. 4, pp. 351–371, 2014.

[358] C. Atzmüller and P. M. Steiner, "Experimental vignette studies in survey research," *Methodology*, 2010.

[359] V. Das Swain, L. Gao, A. Mondal, G. D. Abowd, and M. De Choudhury, "Sensible and sensitive ai for worker wellbeing: Factors that inform adoption and resistance for information workers," in *Proceedings of the CHI Conference on Human Factors in Computing Systems*, 2024, pp. 1–30.

[360] R. Hoyle, L. Stark, Q. Ismail, D. Crandall, A. Kapadia, and D. Anthony, "Privacy norms and preferences for photos posted online," *ACM Transactions on Computer-Human Interaction (TOCHI)*, vol. 27, no. 4, pp. 1–27, 2020.

[361] M. Pinski and A. Benlian, "Ai literacy-towards measuring human competency in artificial intelligence," 2023.

[362] S. Y. Park *et al.*, "Identifying challenges and opportunities in human-ai collaboration in healthcare," in *Companion Publication of the 2019 Conference on Computer Supported Cooperative Work and Social Computing*, 2019, pp. 506–510.

[363] S. T. Mueller *et al.*, "Principles of explanation in human-ai systems," *arXiv preprint arXiv:2102.04972*, 2021.

[364] S. S. Kim, E. A. Watkins, O. Russakovsky, R. Fong, and A. Monroy-Hernández, "" help me help the ai": Understanding how explainability can support human-ai interaction," in *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, 2023, pp. 1–17.

[365] Q. Wang, C. L. Anyi, V. D. Swain, and A. K. Goel, "Navigating ai fallibility: Examining people's reactions and perceptions of ai after encountering personality misrepresentations," *arXiv preprint arXiv:2405.16355*, 2024.

[366] M. Ashoori and J. D. Weisz, "In ai we trust? factors that influence trustworthiness of ai-infused decision-making processes," *arXiv preprint arXiv:1912.02675*, 2019.

[367] D. Wang *et al.*, ""brilliant ai doctor" in rural clinics: Challenges in ai-powered clinical decision support system deployment," in *Proceedings of the 2021 CHI conference on human factors in computing systems*, 2021, pp. 1–18.

[368] S. F. Lyndgaard, R. Storey, and R. Kanfer, "Technological support for lifelong learning: The application of a multilevel, person-centric framework," *Journal of Vocational Behavior*, vol. 153, p. 104 027, 2024.

[369] M. A. DeVito, "Adaptive folk theorization as a path to algorithmic literacy on changing platforms," *Proceedings of the ACM on Human-Computer Interaction*, vol. 5, no. CSCW2, pp. 1–38, 2021.

[370] N. Karizat, D. Delmonaco, M. Eslami, and N. Andalibi, "Algorithmic folk theories and identity: How tiktok users co-produce knowledge of identity and engage in algorithmic resistance," *Proceedings of the ACM on human-computer interaction*, vol. 5, no. CSCW2, pp. 1–44, 2021.

[371] Ş. Sarkadi, A. R. Panisson, R. H. Bordini, P. McBurney, S. Parsons, and M. Chapman, "Modelling deception using theory of mind in multi-agent systems," *AI Communications*, vol. 32, no. 4, pp. 287–302, 2019.

[372] Q. Wang, S. Walsh, M. Si, J. Kephart, J. D. Weisz, and A. K. Goel, "Theory of mind in human-ai interaction," in *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, 2024, pp. 1–6.

[373] C.-M. Huang, S. Andrist, A. Sauppé, and B. Mutlu, "Using gaze patterns to predict task intent in collaboration," *Frontiers in psychology*, vol. 6, p. 1049, 2015.

[374] N. Shaker, S. Asteriadis, G. N. Yannakakis, and K. Karpouzis, "Fusing visual and behavioral cues for modeling user experience in games," *IEEE transactions on cybernetics*, vol. 43, no. 6, pp. 1519–1531, 2013.

[375] W. Samek, T. Wiegand, and K.-R. Müller, "Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models," *arXiv preprint arXiv:1708.08296*, 2017.

[376] G. Guo, E. Karavani, A. Endert, and B. C. Kwon, "Causalvis: Visualizations for causal inference," in *Proceedings of the 2023 CHI conference on human factors in computing systems*, 2023, pp. 1–20.

[377] G. Guo, J. Stasko, and A. Endert, "What we augment when we augment visualizations: A design elicitation study of how we visually express data relationships," in *Proceedings of the 2024 International Conference on Advanced Visual Interfaces*, 2024, pp. 1–5.