Mutual Theory of Mind for Human-AI Communication

Qiaosi Wang[1] , **Ashok K. Goel**[1]
[1]Georgia Institute of Technology
qswang@gatech.edu, ashok.goel@cc.gatech.edu

Abstract

From navigation systems to smart assistants, we communicate with various AI on a daily basis. At the core of such human-AI communication, we convey our understanding of the AI's capability to the AI through utterances with different complexities, and the AI conveys its understanding of our needs and goals to us through system outputs. However, this communication process is prone to failures for two reasons: the AI might have the wrong understanding of the user and the user might have the wrong understanding of the AI. To enhance mutual understanding in human-AI communication, we posit the Mutual Theory of Mind (MToM) framework, inspired by our basic human capability of "Theory of Mind." In this paper, we discuss the motivation of the MToM framework and its three key components that continuously shape the mutual understanding during three stages of human-AI communication. We then describe a case study inspired by the MToM framework to demonstrate the power of MToM framework to guide the design and understanding of human-AI communication.

1 Introduction

As Artificial Intelligence (AI) becomes closely integrated into our daily lives, we communicate with different types of AI systems on a daily basis— we speak to digital assistants to set up reminders and alarms, we receive entertainment recommendations (e.g., music, TV shows) from AI-powered platforms, and sometimes we even rely on AI systems to complete tasks at work. Enhancing the human-AI communication process becomes crucial in many aspects of our lives.

At the core of human-AI communication is a process where the AI system conveys its understanding of the user through system outputs, and the user conveys their understanding of the AI system through feedback. However, this communication process is prone to failure for two reasons: the AI systems might have the wrong understanding of the user and the user might have the wrong understanding of the AI [Riedl, 2019]. While recent advancement in machine learning has made AI systems smarter and better at "reading our minds" by predicting our shopping preferences [Linden *et al.*, 2003] or emotional states [Wang *et al.*, 2020] with fairly high accuracy, errors in machines' interpretations of humans are unavoidable [Riedl, 2019]. Even though much AI systems nowadays can learn from user behaviors to optimize their performance overtime, it is often difficult for the user to tune their mental model of the AI system without clear updates or communication from the AI system [Bansal *et al.*, 2019].

This lack of mutual understanding during communication process is not entirely new, in fact, it is very common at early stages of human-human communication. Yet humans are able to enhance mutual understanding gradually through further communications fairly easily. This is based on a uniquely human characteristic called "Theory of Mind" [Baron-Cohen *et al.*, 1985; Premack and Woodruff, 1978].

Scholars posit that the **Theory of Mind (ToM)** is a basic cognitive and social characteristic that enables us to make conjectures about each others' minds through observable or latent behavioral and verbal cues [Baron-cohen, 1999; Gopnik and Wellman, 1992]. Having the capability of ToM enables us to construe a mental model of others' minds, which includes their thoughts, preferences, goals, needs, plans, etc. [Baron-cohen, 1999; Premack and Woodruff, 1978]. In typical human-human interactions, having a **Mutual Theory of Mind (MToM), meaning all parties involved in the interaction possess the ToM**, enables us to continuously refine our understanding of each others' minds through behavioral and verbal feedback, helping us to maintain constructive and coherent communications. However, we have not been able to accomplish MToM in human-AI communication, which makes mutual understanding in human-AI communication difficult to achieve.

In this paper, we posit MToM as a framework to enhance the mutual understanding in human-AI communication. We argue that *MToM as a framework provides a process and content account of human-AI communication* that emphasizes the iterative mutual shaping of each party's perceptions and feedback through different stages of the communication process. We will first review relevant literature on ToM and human-AI communication, then describe the MToM framework in details. Finally, we will provide a case study that was conducted under the guidance of MToM framework.

2 Related Work

2.1 Theoretical Perspectives of Communication

Communication is commonly defined as *"the process of transmitting information and common understanding from one person to another."* [Lunenburg, 2010] To enhance communication, scholars across different disciplines have offered various perspectives to look at communication.

In communication studies, scholars focus on the effectiveness of communication by looking at the different components at play during the communication process. The classic Shannon-Weaver model of communication [Shannon, 1948] outlines several key components during the communication process [Lunenburg, 2010]: *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. According to this model, the quality and effectiveness of the communication are dependent on these key components during the communication process.

The Cognitive Science perspective of communication highlights the critical role of ToM [Premack and Woodruff, 1978]. Humans' ability of ToM enables us to make suppositions of other's minds through verbal and behavioral cues, which is the foundation of human-human communication [Baron-cohen, 1999; Baron-Cohen *et al.*, 1985]. From this perspective, both interlocutors during communication can form mental models 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 [Premack and Woodruff, 1978; Baron-cohen, 1999]. Based on that mental model we formed about the other's mind, we will act accordingly to correct, explain, or persuade. This cycle of building a mental model of others' minds and then act accordingly continues throughout the communication process. Identifying other's minds through behavioral cues, according to this perspective, is therefore crucial to a smooth and successful communication process.

Communication process can also be interpreted from the social science perspective through impression management [Goffman, 1978]. 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' impressions of them (expression given off)— to manage impressions [Goffman, 1978]. Throughout interactions, each party conveys their definition of the situation through communications, individuals by expressions and audience by reactions to the individuals.

These three theoretical perspectives on communication emphasize different aspects of the communication process: the communication study perspective focuses on the encoding and decoding process of messages; the cognitive science perspective discusses how behavioral cues can inform mental models of interlocutor's minds; the social science perspective describes how perception of others' minds could dictate our behaviors. Our Mutual Theory of Mind framework attempts to bring these different emphasis together into one coherent framework to understand the mutual shaping process of perceptions and feedback during communication.

2.2 Theory of Mind in Human-AI Communication

Over the years, many researchers have recognized the crucial role ToM plays in human-AI interaction. In human-robot teaming research, ToM has been intentionally built in as an individual module of the system architecture to help robots monitor world state as well as the human state [Devin and Alami, 2016], to construct simulation of hypothetical cognitive models of the human partner to account for human behaviors that deviate from original plans [Pynadath and Marsella, 2005], and to help robots to build mental models about user beliefs, plans and goals [Kim and Lipson, 2009; Harbers *et al.*, 2009]. Robots built with ToM have demonstrated positive outcomes in team operations [Devin and Alami, 2016] and are perceived to be more natural and intelligent [Lin *et al.*, 2010].

Other research in human-computer interaction and human-centered AI has also been exploring along the realm of ToM, focusing mostly on enhancing user's mental models and understanding of the AI systems. Prior research has explored people's mental model of AI systems— people's mental model of AI agents could include global behavior, knowledge distribution, and local behavior [Gero *et al.*, 2020]. People's perception of AI systems is instrumental in guiding how they interact with AI systems [Gero *et al.*, 2020] and thus serves as a precursor to their expectation of AI's behavior. Some recent research has also begun to examine how to automatically infer user's mental model of AI. Prior research suggests the potential of leveraging linguistic cues to indicate people's perception of AIs during human-AI interactions. Researchers have been able to infer users' emotions towards an AI agent [Skowron *et al.*, 2011] and signs of conversation breakdowns [Liao *et al.*, 2018]from communication cues.

Given that AI's behavior and output could also influence user's mental model of the AI, and therefore how the user decides to interact with the AI, we want to highlight that the perception-feedback loop is mutual during the human-AI communication process— user's mental model can be informed by AI's behavior yet AI also constructs a mental model of the user, which determines whether or how the AI's output could inform the user about the AI's behavior and working mechanism. We propose the Mutual Theory of Mind framework to capture this mutual shaping of

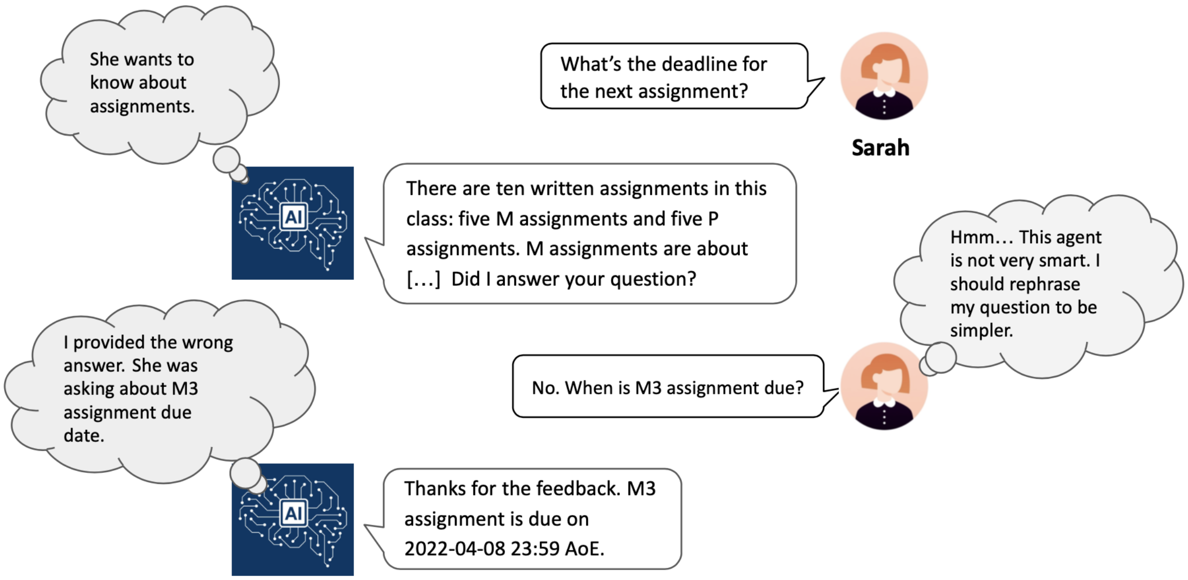

Figure 1: An example human-AI communication dialogue between a student Sarah and a virtual teaching assistant (question-answering agent). This dialogue illustrates the three key elements of MToM in human-AI communication: perception, feedback, and mutuality. Cloud bubble represents *perceptions*, rectangular bubble represents *feedback*, *mutuality* is represented throughout the dialogue in which the AI's recommendation shapes user's perception of AI's capability and the AI's perception of the user is shaped by user's feedback.

perception-feedback during human-AI communication.

3 Mutual Theory of Mind Framework

Drawing from theoretical and empirical research reviewed in the previous section, we posit the MToM framework to guide the understanding and design of enhancing mutual understanding in human-AI communication. The MToM framework provides both process and content account of human-AI communication by highlighting *three elements* that mutually shape the human-AI communication process in *three stages*.

3.1 Three Elements of MToM Framework

In the MToM framework, three elements are critical for humans and AI to reach mutual understanding during the communication process: *perception, feedback, and mutuality*.

In human-AI communication, humans and AI can each construct and revise their *perceptions* of each other based on feedback from the other party. In the MToM framework, this kind of perceptions not only refers to "my understanding about your mind", which is the typical definition of ToM, but also refers to "my understanding of your understanding of my mind", highlighting the recursive property of the perceptions during communication.

Feedback, often in the form of verbal or behavioral cues, is generated with different complexities based on the perceptions of each other. For instance, in a human-chatbot conversation, humans would generate simpler command when they believed the chatbot could not understand complex human language; the chatbot would generate simpler feedback when they interpreted the human needs were simple (e.g., asking about the weather).

While each party involved in the communication is capable of constructing perceptions and generating feedback on their own, communication is a two-way interaction, which means all parties involved in the communication process are *mutually shaping* each other's perceptions through feedback.

These key elements play an important part in determining the success and failure of a communication— mutual understanding can be severely undermined by failure to mutually shape each other's perceptions in the correct way through inaccurate feedback. Figure 1 illustrates these three elements through an example dialogue between the user Sarah and the question-answering virtual teaching assistant in the context of online learning.

3.2 Three Stages of MToM Framework

In the MToM framework, these three elements are constantly shaping the mutual understanding between humans and AI during three stages of the human-AI communication process: *the construction of AI's ToM, the recognition of AI's ToM, and the revision of AI's ToM*.

In the first stage, *the construction of AI's ToM*, the AI system takes in user feedback and tries to interpret what's on the user's mind, specifically, the user's goals, needs, preferences etc. Based on the interpretation, the AI *constructs* its theory of the user's mind, which helps the AI to generate responses accordingly to help the user fulfill their goals and needs in this conversation.

After the AI generates its response to the user, the user then *recognizes* the AI's theory of the user's mind based on AI's response. This recognition leads the user to construct their theory of the AI's mind, which includes how the AI interprets

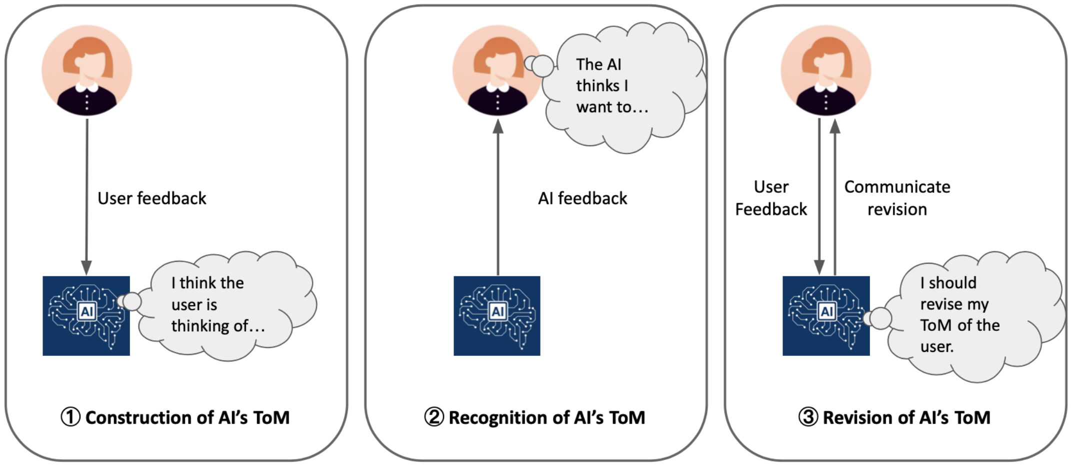

Figure 2: Three stages of human-AI communication in the Mutual Theory of Mind framework.

the user's mind as well as inferences about AI's capability and working mechanism.

The AI's ToM might not always be accurate. Based on the user's mental model of the AI, the user provides feedback/follow-up to the AI, which the AI takes in to *revise* or update its theory of the user's mind based on the user's feedback. In order to enhance mutual understanding between the human and the AI during the communication process, it is crucial for the AI to also communicate this revision of its ToM back to the user through feedback.

We further illustrate the three stages of MToM framework in Figure 2. Throughout these three stages, the three key elements of MToM (perception, feedback, and mutuality) interact with each other to constantly shape the mutual understanding between the human and the AI.

4 MToM Case Study

Guided by the MToM framework, we have been conducting studies to offer empirical insights into how the interplay between perception, feedback, and mutuality shape the mutual understanding between humans and AI throughout the three stages of MToM. In this section, we describe an empirical study inspired by MToM to understand how to automatically construct user's perception of a Conversational Agent (CA) during communication by leveraging the linguistic features from user utterances. This work was published at ACM CHI 2021 conference [Wang *et al.*, 2021].

4.1 Motivation & Context

Conversational Agents (CAs) are becoming increasingly integrated into various aspects of our lives. While CAs are relatively successful in task-oriented interactions, 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 [Clark *et al.*, 2019]. This "gulf" between user expectation and experience with CAs [Luger and Sellen, 2016] has led to constant user frustration, frequent conversation breakdowns, and eventual abandonment of CAs [Luger and Sellen, 2016; Zamora, 2017].

To understand how we could potentially mitigate or bridge this "gulf" between user expectation and experience with the CAs, we took inspiration from the MToM in human-human communication, where our human capability of ToM helps us build a shared expectation of each other through behavioral feedback, enabling us to maintain constructive and coherent conversations. In human-AI communication, if we equipped the CA with an analog of ToM that can automatically identify and construct user perceptions about the CAs through users' communication cues, the CA would be able to monitor users' changing perceptions and provide subtle behavioral cues accordingly to help users build a better mental model about CA's capability. This would also help alleviate the current one-sided communication burden on users, who had to constantly adjust their mental model of the CA through an arbitrary trial-and-error process to elicit desired CA responses [Ashktorab *et al.*, 2019].

This study thus focuses on the first stage of MToM— the construction of AI's ToM, with an emphasis on *constructing AI's theory of the user's perception of the AI through communication cues*. This study took place in an online education context where CAs are prevalent in offering informational and social support to a community of students on online discussion forums. While dyadic interaction between humans and CA have been common across different contexts, in online education environment, CAs face a unique challenge in interacting with both the individual students and the student community as a whole on the discussion forum— each dyadic interaction between individual student and the CA is also visible to the entire student community, therefore changing the entire community's perception of the CA.

4.2 Method & Analysis

To understand how we can construct a community-facing AI's ToM through communication cues, we deployed a community-facing question-answering CA named Jill Watson [Goel and Polepeddi, 2016] (JW for short) in an on-

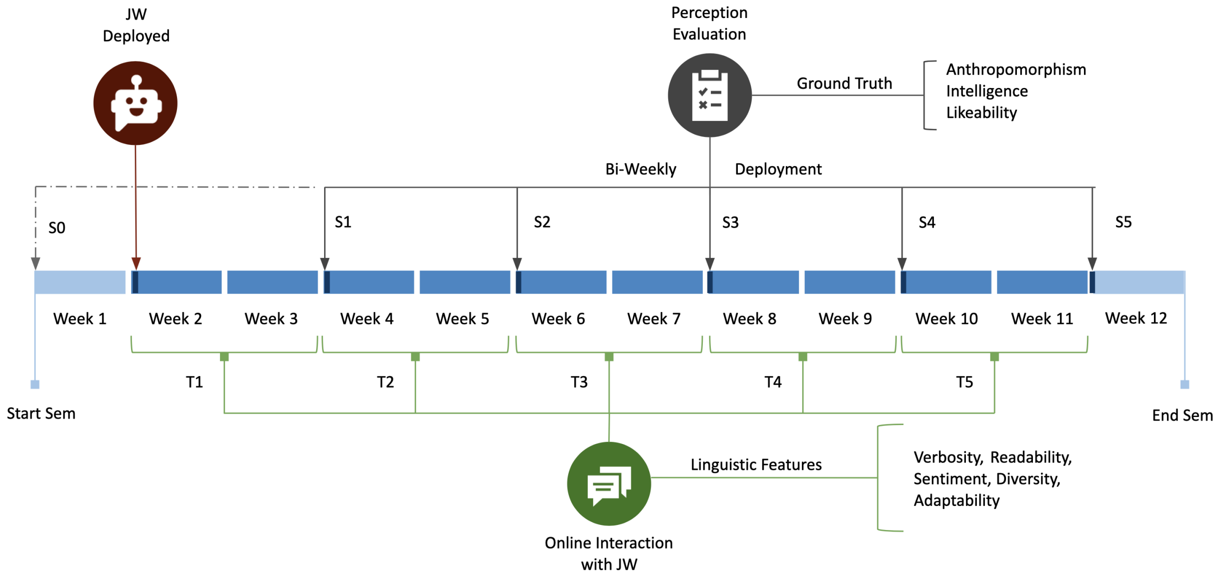

Figure 3: Study design and timeline. S0-S5 represents the survey data. T1-T5 represents our division of class discussion forum data based on the survey distribution timeline. In the regression analysis, we used survey data as ground truth to tag student interaction with JW in each time frame. For instance, we used S1 to tag forum data from T1, S2 to tag T2, and so on.

line class discussion forum to answer students' questions for 10 weeks over the course of a semester. We collected students' bi-weekly self-reported perceptions of JW (anthropomorphism, intelligence, likeability) and conversations with JW for further analysis (see Figure 3 for detailed study design). The perception survey metrics were intentionally selected because they could capture students' holistic social perceptions, and the linguistic features were selected by prior literature to be able to potentially predict these perceptions.

We then built three linear regression models where each model uses one of the three perception measures as the dependent variable, and the linguistic features as independent variables. As both perception and linguistic interactions could be a function of time, we include an ordinal variable of the week of the datapoint as a covariate in the models. Further, we control our 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 1 describes our linear regression models, where $\mathcal{P}$ refers to the measures of anthropomorphism, intelligence, and likeability.

$$\begin{aligned} \mathcal{P} \sim Baseline + Week + Verbosity + Readability \\ + Sentiment + Diversity + Adaptability \end{aligned} \tag{1}$$

Summary of Models

Our linear regression models reveal significance with $\mathtt{R}^2(\mathtt{Anth.}) = 0.85$, $\mathtt{R}^2(\mathtt{Intel.}) = 0.93$, $\mathtt{R}^2(\mathtt{Like.}) = 0.95$; all with $\mathtt{p} < 0.001$. Table 1 summarizes the coefficients of each dependent variable. First, we note the statistical significance of the control variables, *week* and *baseline word use*. We find that people who are more expressive are more likely to have a positive perception of JW on all three perception measures. We find verbosity to be negatively associate with each measure of perception, while adaptability, diversity, and readability positively associate with student perception of JW.

4.3 Findings

Next, we describe our operationalization and observation for each of our linguistic features below.

Verbosity

Drawing on prior work [Saha and Sharma, 2020], we use two measures to describe the verbosity of students' posts: 1) *length* and 2) *linguistic complexity*. We operationalize *length* as the number of unique words per post, and *complexity* as the average length of words per sentence [Saha and Sharma, 2020]. Our regression model (Table 1) suggests that both verbosity attributes show negative coefficients with all the perception measures along with statistical significance. This suggests 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.

Readability

Readability refers to the level of ease readers can comprehend a given text. To capture the readability of students' posts to JW, we 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. Our 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 suggests that readability is a strong predictor of students' perception and positively varies with perception.

Table 1: 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	***

Sentiment

To measure the sentiment of each post to JW, we used the VADER sentiment analysis model [Hutto and Gilbert, 2014], which is a rule-based sentiment analysis model that provides numerical scores ranging from -1 (extreme negative) to +1 (extreme positive). Our regression model (Table 1) 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.

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. We draw on prior work [Saha and Sharma, 2020] to obtain linguistic diversity, and use word embeddings for this purpose. According to our regression model, we find that greater linguistic diversity does not lead to greater perceived intelligence and likeability, whereas, greater linguistic diversity does show a statistically significant positive correlation (0.17) with anthropomorphism.

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 [Goffman, 1978]. Motivated by Saha and Sharma's approach [Saha and Sharma, 2020], we 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. Our regression model indicates that adaptability positively associates with anthropomorphism (1.02), intelligence (1.53), and likeability (2.55), all with statistical significance. Our observations suggest that adaptability is a valid predictor of the perceptions of JW.

Summary and Interpretations.

We examined the relationship between linguistic features of student-JW conversations and student perception of JW through regression analysis. We find that verbosity negatively associates with student perception of JW, whereas readability, sentiment, diversity, and adaptability positively associate with anthropomorphism, intelligence, and likeability. Our findings suggest feasibility for AIs to automatically construct user's perception of the AI through communication cues.

5 Conclusion

The foundation for communication success is for all parties involved in the communication process to reach a mutual understanding of each other. In this paper, we posited the theoretical framework of Mutual Theory of Mind to understand and design human-AI communication. The MToM framework highlighted three key elements— perception, feedback, mutuality— that continuously interact with each other throughout three stages of the communication process— construction, recognition, and revision of AI's Theory of Mind. The MToM framework thus provides a process and content account of human-AI communication that emphasizes on communication as a mutual-shaping process. We further provided a case study that demonstrated an innovative use of feedback and perception under the guidance of the MToM framework to enhance AI's understanding of user's perception of the AI during communication.

References

[Ashktorab *et al.*, 2019] Zahra Ashktorab, Mohit Jain, Q. Vera Liao, and Justin 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*, pages 1–12, New York, New York, USA, 2019. ACM Press.

[Bansal *et al.*, 2019] Gagan Bansal, Besmira Nushi, Ece Kamar, Walter S Lasecki, Daniel S Weld, and Eric Horvitz. Beyond accuracy: The role of mental models in human-ai team performance. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, volume 7, pages 2–11, 2019.

[Baron-Cohen *et al.*, 1985] Simon Baron-Cohen, Alan M Leslie, Uta Frith, et al. Does the autistic child have a "theory of mind". *Cognition*, 21(1):37–46, 1985.

[Baron-cohen, 1999] Simon Baron-cohen. Evolution of a Theory of Mind? In *The Descent of Mind: Psychological Perspectives on Hominid Evolution*, pages 1–31. Oxford University Press, 1999.

[Clark *et al.*, 2019] Leigh Clark, Nadia Pantidi, Orla Cooney, Philip Doyle, Diego Garaialde, Justin Edwards, Brendan Spillane, Emer Gilmartin, Christine Murad, Cosmin Munteanu, Vincent Wade, and Benjamin R. Cowan. What makes a good conversation? Challenges in designing truly conversational agents. *Conference on Human Factors in Computing Systems - Proceedings*, pages 1–12, 2019.

[Devin and Alami, 2016] Sandra Devin and Rachid Alami. An implemented theory of mind to improve human-robot shared plans execution. *ACM/IEEE International Conference on Human-Robot Interaction*, 2016-April:319–326, 2016.

[Gero *et al.*, 2020] Katy Ilonka Gero, Zahra Ashktorab, Casey Dugan, Qian Pan, James Johnson, Werner Geyer, Maria Ruiz, Sarah Miller, David R. Millen, Murray Campbell, Sadhana Kumaravel, and Wei Zhang. Mental Models of AI Agents in a Cooperative Game Setting. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–12, New York, NY, USA, 4 2020. ACM.

[Goel and Polepeddi, 2016] Ashok K Goel and Lalith Polepeddi. Jill watson: A virtual teaching assistant for online education. Technical report, Georgia Institute of Technology, 2016.

[Goffman, 1978] Erving Goffman. *The Presentation of Self in Everyday Life*. London: Harmondsworth, 1978.

[Gopnik and Wellman, 1992] Alison Gopnik and Henry M Wellman. Why the child's theory of mind really is a theory. 1992.

[Harbers *et al.*, 2009] Maaike Harbers, Karel Van Den Bosch, and John Jules Meyer. Modeling agents with a theory of mind. *Proceedings - 2009 IEEE/WIC/ACM International Conference on Intelligent Agent Technology, IAT 2009*, 2:217–224, 2009.

[Hutto and Gilbert, 2014] 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.

[Kim and Lipson, 2009] Kyung-Joong Kim and Hod Lipson. Towards a simple robotic theory of mind. page 131, 2009.

[Liao *et al.*, 2018] Q. Vera Liao, Werner Geyer, Muhammed Mas-ud Hussain, Praveen Chandar, Matthew Davis, Yasaman Khazaeni, Marco Patricio Crasso, Dakuo Wang, Michael Muller, and N. Sadat Shami. 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*, volume 8, pages 1–13, New York, New York, USA, 2018. ACM Press.

[Lin *et al.*, 2010] Shuhong Lin, Boaz Keysar, and Nicholas Epley. Reflexively mindblind: Using theory of mind to interpret behavior requires effortful attention. *Journal of Experimental Social Psychology*, 46(3):551–556, 2010.

[Linden *et al.*, 2003] Greg Linden, Brent Smith, and Jeremy York. Amazon. com recommendations: Item-to-item collaborative filtering. *IEEE Internet computing*, 7(1):76–80, 2003.

[Luger and Sellen, 2016] Ewa Luger and Abigail 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*, pages 5286–5297, 2016.

[Lunenburg, 2010] Fred C Lunenburg. Communication: The process, barriers, and improving effectiveness. *Schooling*, 1(1):1–10, 2010.

[Premack and Woodruff, 1978] David Premack and Guy Woodruff. Does the chimpanzee have a theory of mind? *Behavioral and brain sciences*, 1(4):515–526, 1978.

[Pynadath and Marsella, 2005] David V. Pynadath and Stacy C. Marsella. PsychSim: Modeling theory of mind with decision-theoretic agents. *IJCAI International Joint Conference on Artificial Intelligence*, pages 1181–1186, 2005.

[Riedl, 2019] Mark O Riedl. Human-centered artificial intelligence and machine learning. *Human Behavior and Emerging Technologies*, 1(1):33–36, 2019.

[Saha and Sharma, 2020] Koustuv Saha and Amit Sharma. Causal Factors of Effective Psychosocial Outcomes in Online Mental Health Communities. In *ICWSM*, 2020.

[Shannon, 1948] Claude Elwood Shannon. A mathematical theory of communication. *The Bell system technical journal*, 27(3):379–423, 1948.

[Skowron *et al.*, 2011] Marcin Skowron, Stefan Rank, Mathias Theunis, and Julian Sienkiewicz. The good, the bad and the neutral: affective profile in dialog system-user communication. In *International Conference on Affective Computing and Intelligent Interaction*, pages 337–346. Springer, 2011.

[Wang *et al.*, 2020] Qiaosi Wang, Shan Jing, David Joyner, Lauren Wilcox, Hong Li, Thomas Plötz, and Betsy Disalvo. 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*, pages 63–76, 2020.

[Wang *et al.*, 2021] Qiaosi Wang, Koustuv Saha, Eric Gregori, David Joyner, and Ashok 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*, pages 1–14, 2021.

[Zamora, 2017] Jennifer 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*, pages 253–260, 2017.