Keywords – Human-Computer Interaction, Social Computing, AI-Mediated Communication, Social Cognition

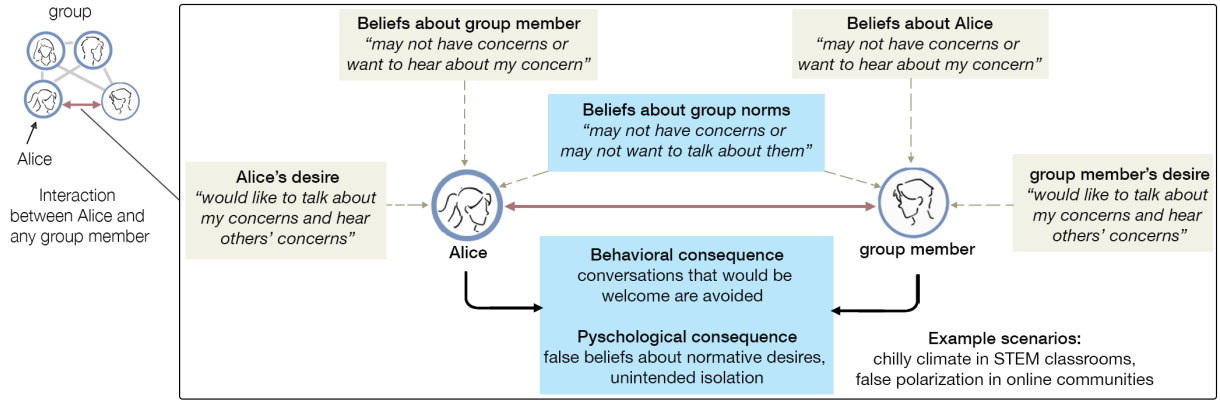
Within the field of human-computer interaction (HCI), a long-standing approach to designing social technologies for prosocial outcomes relies on reinforcement: using rewards and punishments to encourage certain observable behaviors (e.g., badges for participating in a discussion). By treating people as "black boxes" who produce behaviors in response to stimuli, this approach falls short when the goal is not simply to produce a behavioral outcome but also a psychological one (e.g., reducing partisan animosity, making people feel safer, making people feel more included). Our systems need to appropriately account for the causes of observable behaviors (e.g., are expressions of assent by group members the result of agreement or reluctance to address conflict?). To take on this challenge, my research in HCI looks inside the "black box": focusing design efforts not only on observable behaviors, but also on the interpretation and sensemaking processes people bring to social interaction. Further, my work shows how explicitly accounting for these mental processes is broadly useful in computing research: it not only helps us realize prosocial outcomes but also improves our understanding of how people interact with and through AI systems.

In social interactions, people employ several cognitive tools, such as perspective-taking, categorizing others as "us" and "them", making inferences about others' mental states, and explaining people's behavior. By examining these tools, studied in psychology as *social cognition*, my work aims to make progress on two interrelated topics:

- Developing **novel systems for mediated social interactions** that improve local and civic participation
- Developing **theoretical frameworks** to understand how AI systems reshape social interactions, as they increasingly mediate our interactions and also take on social agentic roles

To pursue this work, I combine insights from social cognition with technical, design, and experimental approaches. Where possible, I deploy systems in the public to achieve real-world impact, and to study their use. My systems have been used by CMU REU students to connect with each other, by small business owners in Pittsburgh to collectively navigate their entrepreneurial journeys, and by open source software users to provide positive feedback to contributors. While my core strengths lie in HCI, my work is enriched through collaborations across disciplines. I have collaborated with researchers in natural language processing (NLP), psychology, and behavioral economics. This research program has yielded publications at top-tier HCI venues such as CHI and CSCW, and has been recognized with two Best Paper Honorable Mention awards. To support this work, so far, I have raised funding (totaling $210,000) from NSF and Google. Next, I summarize my past research on each topic, highlighting key ideas and contributions, before describing planned future directions.

Improving local and civic participation

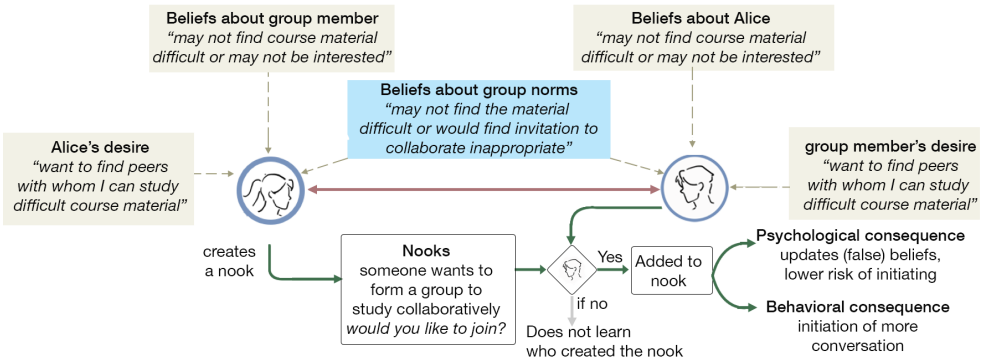


An illustration of social deadlocks that can impede local and civic participation. For simplicity, the illustration shows one dyad. Actually, of course, any member could be in Alice's position. Each member of the group has three roles in the interaction: first, each member is a potential conversation initiator (Alice); second, each member can be the group member whom the initiator considers approaching, who may or may not welcome the interaction; third, each member contributes to the perceived group norms.

People in similar situations—such as students taking a course together, or small business owners in a locality—who may want to connect and navigate their situations collectively, often fail to do so. For example, a student looking

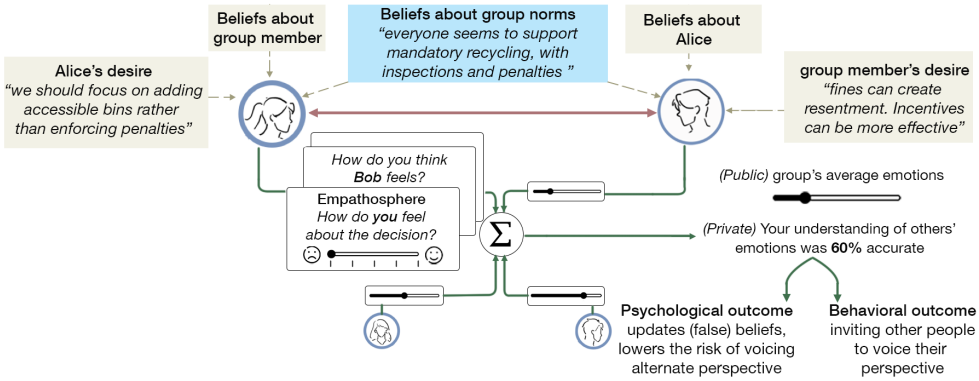
to initiate interactions may worry (justifiably or not) about others being disinterested. Even when groups begin interacting, people may avoid speaking up out of concern for how their views will be received. Misplaced psychological barriers often arise because we inaccurately assess others' views before interacting with them, a result of our egocentric perspective-taking abilities [Epley et al., 2022]. Because we tend to make inferences about others' minds by using our own minds as a guide, we often form miscalibrated beliefs about how others will respond. For instance, a person starting a conversation with a stranger might worry that the success of the conversation depends on how effectively they can start and maintain the conversation, whereas the stranger's experience is determined more by the friendliness conveyed by starting the conversation. This creates social deadlock: many may want to connect or hear dissenting views, but few are willing to take the first step. Overcoming these deadlocks is crucial for fostering spaces where people feel they belong and can engage in constructive dialogue. The first contribution of my work is a set of systems that give users the tools for perspective-*getting*: users can account for others' perspectives in the process of taking an action.

Nooks is a system to lower hesitations in interacting with new people in shared digital environments, such as people who may be connected through a Slack workspace, Discord server, or neighborhood groups on Facebook [Bali et al., 2023]. Its insight is that people would feel more comfortable starting conversations if they could know in advance that the people they were about to interact with were interested in talking about the same topic. To test this idea, I led an undergraduate student in developing Nooks as a Slack application. To initiate a conversation, individuals anonymously create a "nook"—a conversation room with a defined topic and norms, but without revealing its creator. The system then probes others in the workspace for a few hours to gauge interest in joining. Once enough interest is shown, the nook becomes active, and those interested are added to the channel. At this point, the creator and interested participants are deanonymized, but all participants share a mutual interest in the topic, reducing the initiator's risk of social evaluation. Collaborating with CMU's REU program, we deployed Nooks for 9 weeks in the students' Slack workspace. We found it provided students with non-threatening and inclusive interaction opportunities, ambient awareness about others interests, and led to new interactions online and offline. This work received a Best Paper Honorable Mention award at ACM CHI 2023.



Top: Uncertainty or false beliefs about how people might react can impede welcome conversations.
Bottom: Nooks lowers the risk of initiating interactions.

Even after groups begin interacting, psychological barriers persist. In public spaces like classrooms, online communities, and citizen assemblies, people may welcome diverging views but hesitate themselves to voice disagreement, fearing social repercussions. As a result, discourse can be dominated by those less vulnerable to these risks, distorting perceived norms and excluding valuable perspectives. Reinforcement-based solutions, like providing feedback on observable behaviors (e.g., who has spoken so far or verbally agreed), can backfire. One system visualizing verbal agreement led members to verbally agree with the majority opinion, even if they disagreed, to improve displayed agreement [Leshed et al., 2009]. *Empathosphere*



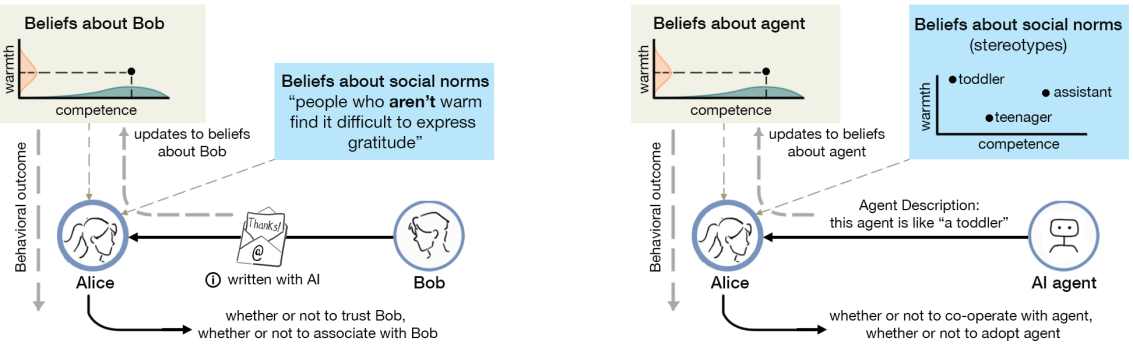
Top: Miscalibrated beliefs about group norms can lead to false polarization and unintended exclusion.
Bottom: Empathosphere provides public and private feedback to support constructive communication.

[Khadpe et al., 2022] presents aggregated perspectives in group settings, ensuring no individual's view can be traced, creating a layer of anonymity that encourages authentic expression. In contrast to prior approaches that reveal patterns in *observable activity* using *incidentally* generated traces, Empathosphere investigates the potential of revealing *latent perspectives* by *explicitly eliciting* them. I conducted a controlled study involving virtual citizen assemblies that compared Empathosphere to a reflection-based intervention. Groups of 4–6 participants worked synchronously on a participatory budgeting task via a chat system. We found that Empathosphere improved satisfaction with group outcomes, encouraged open communication and feedback, and increased the groups' desire to continue working together.

Taken together, Nooks and Empathosphere show that by examining barriers that stem from social cognition processes, we can design new mediums that help realize prosocial outcomes.

Systematizing how people interact through and with AI systems

Looking at the cognitive tools people bring to social interaction is also useful for the inverse goal: making sense of how people interact through new digital mediums and with new social agents.



Left: How do messages written with AI affect beliefs about the sender? **Right**: How does the design of an agent affect the user's (social) evaluations of it?

For instance, a recurring observation in recent empirical work is that if a message is known to be written with AI, the recipient judges the sender as less *warm*— less trustworthy and less kind—than in absence of such knowledge. Despite this frequent observation, few mechanisms have been proposed for why this is the case. A common speculation is that people have an aversion to the use of AI in interpersonal communication, although there remains a lack of theoretical consensus. In my recent work [Khadpe et al., 2024], we question whether AI aversion is the main cause of the effects reported in prior studies. While prior work examined categories of communication that normatively evidence warmth (e.g., condolences, icebreakers), our investigation includes communication that normatively evidences a lack of warmth: brags and blames. Our experiment demonstrates that indication of AI-assistance leads to lower warmth judgments when thanking and apologizing, but not when bragging and blaming. We argue (and our study shows) that diminished perceptions of warmth from prior studies did not occur due to a categorical aversion to AI, but because messages written with AI are viewed as weaker signals of the sender's warmth.

In other work, I've shown how understanding impression formation and stereotyping processes can help explain why Microsoft's chatbot Tay was discontinued for eliciting anti-social troll interactions but Microsoft's Xiaoice, with the same underlying technologies, amassed millions of monthly users [Khadpe et al., 2020]. Specifically, we studied an important and unexamined difference between these otherwise similar agents: the metaphors attached to them. While Tay was presented as "Microsoft's AI fam from the internet that's got zero chill!", signaling high competence and low warmth, Xiaoice was setup to be a "Sympathetic ear", signaling high warmth and even priming behaviors around warmth such as personal disclosure. Our study found that people were more likely to cooperate with a bot with metaphors projecting higher warmth—a result consistent with the fact that Xiaoice continued to remain popular with its user base while Tay was removed within 16 hours of its release for attracting trolls. This work received a Best Paper Honorable Mention award at ACM CSCW 2020.

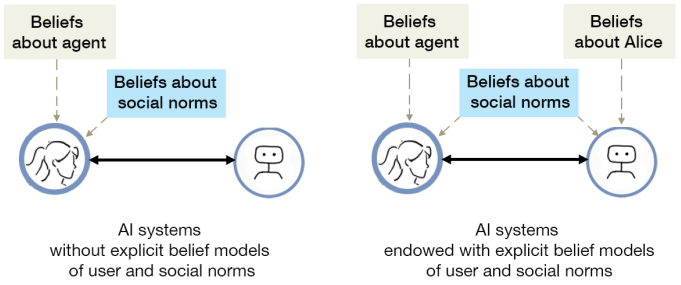
Future research

My goal is to develop digital environments that help people reconnect with each other and with civic life. At the same time, I want to develop theoretical accounts that can guide the design of computational systems introduced in social interaction. Concretely, I envision pursuing the following lines of research in the next 3–5 years.

Self-correcting digital environments. Systems like Nooks and Empathosphere help challenge dominant norms in digital environments. When norms are unclear, such as whether a conversation or perspective is welcome, these systems encourage users to "test the waters" instead of staying silent. However, they are useful only if someone thinks to use them. Could future systems trigger interventions automatically by inferring when someone is likely feeling excluded? I am interested in investigating how we can model individuals' likely internal states (e.g., beliefs) in conversation so that digital environments can intervene adaptively, predicting not just opportune moments to intervene but also the content of those interventions (e.g., a version of Empathosphere that targets specific contested points instead of overall emotional states). Ultimately, I want to design digital environments that are *self-correcting*, with built-in mechanisms to counteract false polarization and unintended exclusion.

Design patterns to overcome social psychological barriers. While my work, and that of others, has applied social cognition theories to design for prosocial outcomes, we are still far from a generalizable framework. How do we systematically close the gap between social scientific knowledge and concrete system designs? One approach I plan to pursue is design patterns—formalizing psychologically informed solutions to recurring problems. For example, Nooks uses *contingent actions* ("start a conversation with them on the condition that they also want to talk") to resolve social deadlocks. Once codified, the pattern can be noticed and applied elsewhere, such as using contingent actions to address deadlocks in collective action ("commit on the condition that others also commit").

Socially-situated design of AI. Modeling the social cognitive processes that activate when people interact with AI systems also reveals new ways to improve human-AI interaction. For instance, existing NLP systems predict what to say—why not also predict how the user might react? In prior work, I've shown how real-world conversational agents can be more successful if they incorporate models of their human partners and act optimally with respect to these models [Bawa et al., 2020]. I am excited about collaborating with NLP researchers to incorporate explicit models of human social cognition in the design of language systems. Alongside this, I am interested in developing measurement toolkits that allow designers, researchers, and policymakers to reason about the activation of specific social cognitive processes when people interact with a particular AI system. This was the focus of a recent proposal I submitted (as Co-PI) to CMU's Block Center for Technology and Society.



We can improve Human-AI Interaction by endowing today's AI systems (**left**) with learnable belief models of their users and social norms (**right**). Figure uses paradigm extended in [Collins et al., 2024].

References

[Khadpe et al., 2024] **Pranav Khadpe**[*], Kimi Wenzel[*], George Loewenstein, Chinmay Kulkarni, Geoff Kaufman. AI-Mediated Communication Revisited: Whether or Not Perceived AI Use Leads to Lower Warmth Judgments Depends on Message Type. *in submission.*

[Bali et al., 2023] Shreya Bali, **Pranav Khadpe**, Geoff Kaufman, Chinmay Kulkarni. Nooks: Social Spaces to Lower Hesitations in Interacting with New People at Work. In *Conference on Human Factors in Computing Systems (**CHI**) 2023.* **Honorable Mention Award**.

[Khadpe et al., 2022] **Pranav Khadpe**, Chinmay Kulkarni, Geoff Kaufman. Empathosphere: Promoting Constructive Communication in Ad-Hoc Virtual Teams through Perspective-Taking Spaces. In *Proceedings of the ACM on Human-Computer Interaction (**CSCW**) 2022.*

[Khadpe et al., 2020] **Pranav Khadpe**, Ranjay Krishna, Li Fei-Fei, Jeffrey Hancock, and Michael Bernstein. Conceptual Metaphors Impact Perceptions of Human-AI Collaboration. In *Proceedings of the ACM on Human-Computer Interaction (**CSCW**) 2020.* **Honorable Mention Award**

[Bawa et al., 2020] Anshul Bawa, **Pranav Khadpe**, Pratik Joshi, Kalika Bali, Monojit Choudhury. Do Multilingual Users Prefer Chat-bots that Code-mix? Let's Nudge and Find Out! In *Proceedings of the ACM on Human-Computer Interaction (**CSCW**) 2020.*

[Epley et al., 2022] Nicholas Epley, Michael Kardas, Xuan Zhao, Stav Atir, Juliana Schroeder. Undersociality: Miscalibrated social cognition can inhibit social connection. In *Trends in Cognitive Sciences 2022.*

[Collins et al., 2024] Katherine M Collins, Ilia Sucholutsky, Umang Bhatt, Kartik Chandra, Lionel Wong, Mina Lee, Cedegao E Zhang, Tan Zhi-Xuan, Mark Ho, Vikash Mansinghka, Adrian Weller, Joshua B Tenenbaum, Thomas L Griffiths. Building machines that learn and think with people. In *Nature Human Behaviour 2024.*

[Leshed et al., 2009] Gilly Leshed, Diego Perez, Jeffrey T Hancock, Dan Cosley, Jeremy Birnholtz, Soyoung Lee, Poppy L McLeod, Geri Gay. Visualizing real-time language-based feedback on teamwork behavior in computer-mediated groups. In *Conference on Human Factors in Computing Systems (**CHI**) 2009.*