WE SHAPE AI, AND THEREAFTER AI SHAPE US: HUMANS ALIGN WITH AI THROUGH SOCIAL INFLUENCES

Jingshu Li
Department of Computer Science
National University of Singapore
21 Lower Kent Ridge Rd, Singapore 119077
jingshu@u.nus.edu

Beichen Xue
Department of Statistics and Data Science
National University of Singapore
21 Lower Kent Ridge Rd, Singapore 119077
xue.beichen@u.nus.edu

Tianqi Song
Department of Computer Science
National University of Singapore
21 Lower Kent Ridge Rd, Singapore 119077
tianqi_song@u.nus.edu

Yi-Chieh Lee
Department of Computer Science
National University of Singapore
21 Lower Kent Ridge Rd, Singapore 119077
yclee@nus.edu.sg

ABSTRACT

The rapid advancement of AI technologies has facilitated the integration of AI into various social scenarios in daily life, making it crucial to understand how AI influences humans and fosters human-AI alignment. While previous research has extensively explored how AI can be consciously designed to affect human behavior, a critical yet underexplored area is how AI can subtly shape human cognition, emotions, and behavior, mirroring *social influence* in human interactions. By examining social influence mechanisms and case studies from recent human-AI interaction studies, this paper identifies two key mechanisms through which AI influences humans: *contagion* and *conformity*. We further explore the challenges and opportunities of AI-driven social influence, urging future research to address risks such as cognitive exploitation and group manipulation, while also leveraging its potential benefits, including fostering positive emotional contagion and supporting healthy behavioral interventions. We advocate for a governance framework that integrates technological, ethical, and societal considerations and call for multidisciplinary collaboration to explore new paradigms of social influence in human-AI interactions. Ultimately, this perspective study provides both theoretical insights and practical pathways toward a harmonious and symbiotic human-AI society.

1 INTRODUCTION

The breakthrough development of artificial intelligence (AI) technologies has profoundly reshaped the paradigms of human-computer interaction (HCI) and human-computer collaboration (Amershi et al., 2019). Intelligent systems represented by generative AI (Achiam et al., 2023; Liu et al., 2024a) have already permeated key social domains such as education, healthcare, and creative design (Biswas, 2023; Zhang et al., 2025; Shaer et al., 2024). As AI increasingly takes on advanced tasks like decision support and knowledge production, the question of how to ensure that AI systems adhere to human values and social ethical norms—often referred to as "AI alignment"—has become a central concern for researchers and practitioners (Goyal et al., 2024; Ouyang et al., 2022). However, human values and ethical expectations evolve, especially as AI integrates into decision-making, col-

laborative work, and social governance. This dynamic interaction leads to mutual adaptation, where humans also refine their perceptions and behaviors in response to AI (D'Amato, 2024). Against this backdrop, the concept of "bidirectional human-AI alignment" has emerged. Its core proposition covers two dimensions: on one hand, AI systems must align with human values; on the other hand, humans must proactively adapt their cognition and interaction strategies to accommodate the behavioral patterns and capability boundaries of intelligent systems (Shen et al., 2024).

Recent HCI research has increasingly focused on "social influence" exerted by AI systems, a concept rooted in social psychology, where individuals adjust their viewpoints, beliefs, or behaviors based on interactions with others (Moussaïd et al., 2013). In human-AI interaction, studies have revealed that humans can also be subject to social influence in their interactions with AI, leading them to align with AI in cognitive, emotional, and behavioral aspects (Li et al., 2025; Song et al., 2024; Shen & Wang, 2023; Herrando & Constantinides, 2021). This represents a shift from earlier research, which primarily examined conscious interactions, such as following AI recommendations or adjusting decisions based on AI-generated feedback. Instead, recent findings reveal that AI can subtly influence human behavior without deliberate effort, facilitating passive adaptation rather than intentional decision-making (Li et al., 2025; Song et al., 2024; Shen & Wang, 2023; Herrando & Constantinides, 2021). This passive adaptation to AI differs from intentional decision-making and highlights a more subconscious process of social alignment with AI-generated inputs.

Within the study of social influence in human-AI interaction, researchers have identified two common forms of influence. The first is "contagion" (Prinz, 2022; Liu et al., 2024b; Tsvetkova et al., 2024; Lee et al., 2020), as seen, for example, in customer-service chatbots whose positive emotional expressions can induce more positive emotions in users (Prinz, 2022). The second is "conformity" (Köbis et al., 2021; Brandstetter et al., 2014; Vollmer et al., 2018; Hertz & Wiese, 2018), in which people may abandon their original positions—despite holding different opinions—when AI provides a specific suggestion or judgment, owing to AI's perceived "authority" or the social pressure it generates, thus exhibiting conformity in group decision-making or collaboration (Hertz & Wiese, 2018).

In part, these social influence phenomena can be explained by the "Computers Are Social Actors" (CASA) theoretical framework (Nass et al., 1994; Gambino et al., 2020). According to CASA, when interacting with computers or AI systems, humans tend to perceive them as "human-like" agents endowed with social attributes, thereby displaying response patterns similar to those shown in human-to-human interaction (Gambino et al., 2020).

As AI becomes increasingly anthropomorphic, more intelligent, and equipped with stronger decision-making and computational capabilities, its influence in social interactions expands significantly (Gong, 2008). This intensifies the urgency of addressing human-to-AI alignment resulting from social influence. If such alignment happens without informed consent or voluntary participation, it may become a form of "exploitation" of individuals' cognition and behavior, leading to personal harm by inducing or manipulating irrational changes, and raising the possibility of "cognitive warfare" or political manipulation on a social scale (Li et al., 2025; Tsvetkova et al., 2024; Köbis et al., 2021). Conversely, if managed and leveraged properly, it could produce positive social outcomes by, for instance, promoting positive emotions, or enabling health behavior interventions (Pescetelli & Yeung, 2022; Li et al., 2025). Therefore, an in-depth exploration of human-to-AI alignment driven by social influence is crucial for building a future harmonious human-AI society (Floridi et al., 2018) and fostering human-AI symbiosis (Jarrahi, 2018), while also offering a key perspective for understanding and designing the next generation of AI collaboration systems.

In the following sections, we will focus on the two representative types of social influence—contagion and conformity—accompanied by illustrative examples. We will then further explore the potential risks and opportunities arising from these social influences of AI, and propose a possible future research agenda. On this basis, we call for greater attention from academia and industry to deepen research and collaboration in this area.

2 CONTAGION AND CONFORMITY: SOCIAL INFLUENCES FROM AI

2.1 CONTAGION

By sociological theory, *contagion* is defined as the process by which behaviors, emotions, or opinions spread among individuals in a non-coercive manner (Wheeler, 1966; Hatfield et al., 1993; Christakis & Fowler, 2013). When individuals perceive others' expressions, actions, or language, they may unconsciously replicate these patterns (Chartrand & Bargh, 1999). Some neuroscience studies suggest that this phenomenon may stem from automatic mirroring and synchronization mechanisms within the human mirror neuron system (Prochazkova & Kret, 2017; Gallese, 2009). Many contagion processes occur at a subconscious level and can be activated through multimodal interactions such as text or speech (Dimberg et al., 2000; Gallese, 2009). With the widespread adoption of socialbots and generative AI, researchers have found that the phenomenon of contagion not only occurs in interpersonal interactions but can also be observed in human-AI interaction.

One prominent form of contagion in human-AI interactions is emotional contagion, where AI-expressed emotions influence human emotions through affective pathways. Research shows that AI expressing positive emotions can enhance customer experiences and enrich service interactions by transmitting positive affect to users (Han et al., 2023; Prinz, 2022). Similarly, AI systems that demonstrate empathic concern and emotional mimicry can increase users' arousal and pleasure, leading to greater engagement and a higher likelihood of continued interaction (Liu et al., 2024b).

Beyond emotional contagion, researchers have also identified *metacognitive* contagion in human-AI interactions. A recent study (Li et al., 2025) revealed that in human-AI decision making, individuals' self-confidence aligns with the level of uncertainty expressed by the AI through confidence scores. This phenomenon influences the calibration between confidence and accuracy, thereby affecting the efficacy of human-AI decision making (Li et al., 2025).

AI systems can also shape human through social and behavioral contagion. AI-driven agents play a significant role in social contagion within online environments, particularly in opinion formation and public discourse. A model-based simulation study has shown that manipulative AI actors and social bots can amplify specific narratives, shift public sentiment, and even contribute to the spiral of silence—a phenomenon in which individuals refrain from expressing dissenting opinions due to perceived social pressure (Ross et al., 2019).

These findings illustrate that contagion effects in human-AI interactions mirror many of the dynamics observed in human-human interactions, raising important questions about how AI-mediated contagion influences trust, emotions, decision-making, and collective behavior in digital and social environments.

2.2 CONFORMITY

Conformity is a fundamental social phenomenon in which individuals adjust their thoughts, behaviors, or decisions to align with a group or prevailing social norms (Asch, 1956). Unlike social imitation, which involves replicating others' actions without external pressure, conformity occurs due to social pressure, compelling individuals to change their behaviors even when they might privately disagree. This psychological mechanism has been widely studied in human-human interactions, demonstrating how peer influence can shape individual decision-making. Conformity is typically classified into normative and informational conformity (Deutsch & Gerard, 1955). Normative conformity occurs when individuals adjust their behavior to gain social acceptance or avoid rejection, even if they do not internally agree with the majority. Informational conformity, on the other hand, happens when individuals accept external information as a reflection of reality, especially in uncertain or ambiguous situations.

As AI systems become increasingly integrated into everyday life, research has shown that humans also conform to AI agents, raising important questions about the nature and implications of such interactions. Recent studies on human-AI interactions reveal that both types of conformity emerge when people interact with AI agents, with stronger effects observed as the number of AI agents increases (Song et al., 2024). These findings suggest that AI, much like human groups, can exert social influence, shaping users' behaviors and decisions in both explicit and subtle ways. Informational conformity is particularly evident when AI provides additional information, leading participants

to align their decisions with AI recommendations (Salomons et al., 2021; Masjutin et al., 2022), whereas normative conformity emerges when individuals are aware they hold a minority position relative to AI agents (Salomons et al., 2021).

These studies also identify key factors and contextual influences that shape conformity in human-AI interactions. One crucial factor is the nature of the task—in objective decision-making tasks, individuals are more likely to conform to AI-generated responses, whereas in subjective tasks, they tend to align more with human judgments (Riva et al., 2022). Trust also plays a significant role in AI conformity. While initial trust in AI can lead to conformity levels comparable to those observed in classic human conformity studies, repeated AI errors diminish trust, making users less likely to follow AI recommendations (Salomons et al., 2018; Hertz & Wiese, 2018). Moreover, age differences influence AI conformity effects. Studies show that children are more susceptible to AI-induced social pressure, whereas adults are more resistant, suggesting developmental differences in how people respond to AI influence (Vollmer et al., 2018). Beyond conventional AI interfaces, virtual and digital AI agents—such as avatars in immersive environments—can exert social and moral pressure, much like human groups, influencing individuals' behaviors and judgments (Bocian et al., 2024; Kyrlitsias et al., 2020).

These findings highlight the complex interplay between AI presence, trust, task characteristics, and user demographics in shaping human conformity in AI interactions, emphasizing the need for further exploration of how AI influence can be moderated or leveraged in different contexts.

3 Challenges, Opportunities and Future Work

In exploring human-AI bidirectional alignment, the social influence exerted by AI on humans poses significant challenges while also offering critical opportunities. Understanding and managing these influences is thus central to shaping the social order in a future human-machine society.

A prominent challenge lies in how AI amplifies the contagion of emotions and viewpoints within social networks. When AI, through recommendation algorithms or automated content generation, shapes users' reading experiences and interaction environments, negative emotions or extreme views can spread with greater efficiency (Tsvetkova et al., 2024; Kirk et al., 2025). Without effective monitoring and intervention, group anxiety, polarization, and rumor propagation can escalate, heightening the risk of social fragmentation and conflict. For example, in decision-making contexts, AI-driven social pressure may reinforce conformity, leading to potentially severe consequences for group decisions. (Liel & Zalmanson, 2020; Riva et al., 2022). When AI is perceived as an "expert" or an "authority," people tend to assume that its judgments are correct and may overlook minority opinions or self-doubt in group deliberations, thus reducing oversight and critical reflection in the decision-making process. If AI's recommendations contain systemic biases or errors that users adopt on a large scale under the influence of social pressure or conformity, the negative impact may escalate from the individual to the systemic level, causing widespread risk (Gabriel, 2020; Köbis et al., 2021).

Even more concerning is that AI with biased values can gradually assimilate social groups, as it reinforces biases and creates a self-perpetuating cycle of radicalization that threatens diversity and stability. As biased outputs feed back into training data, AI-generated content becomes increasingly extreme, shaping public perception over time. Beyond this unintended influence, AI could also serve as a powerful tool for commercial and political interests through deliberate manipulation, subtly reshaping individual and collective values from the theory of *spectacle society* (Debord, 2021).

Nevertheless, the social influence of AI also presents opportunities for human-machine society. On the one hand, if the values and social norms carried by AI are positive and inclusive, contagion and consensus-building processes can generate a virtuous cycle of group cohesion (Tomašev et al., 2020). Leveraging AI to spread beneficial emotions and values may lead to more harmonious social functioning and support further advancement on the path of human-AI collaboration and symbiosis (Pedreschi et al., 2024; Jarrahi, 2018). On the other hand, social influence can be deliberately harnessed for positive goals, such as behavioral correction or social welfare (Oliveira et al., 2021; Li et al., 2025). In the health domain, for example, AI can employ gentle emotional reminders and social comparisons to help people break bad habits or maintain fitness programs. In education, AI may serve as a personalized behavioral role model, promoting more effective study habits aligned with each learner's style and progress. In these scenarios, AI leverages appropriate social influ-

ence mechanisms to support public well-being in areas such as health, education, and environmental protection, demonstrating the potential synergy between technology and social governance.

For future research, the first step is to expand the study of social influence mechanisms, moving beyond contagion and conformity to explore other influence modalities and their variations across cultural and societal contexts. Accurately capturing how AI affects human psychology and behavior across diverse interactive settings requires interdisciplinary collaboration involving psychology, sociology, computer science, and HCI, among others. Second, it is essential to identify and quantify various factors that shape social influence, including individual differences, AI representations, interaction design, and social environments, and to investigate their interactive effects. On this foundation, building responsible AI research and deployment frameworks also becomes vital. This includes implementing traceability, auditability, and explainability mechanisms in algorithm training, data annotation, and content distribution processes. Only through collective efforts to regulate AI development and application can we minimize negative outcomes while maximizing the social benefits AI can offer.

In sum, AI's role in shaping future societies inevitably involves multi-layered social influence on humans. A deeper, multidisciplinary understanding of these challenges and opportunities is crucial for developing effective design strategies and governance measures that promote a healthier, more sustainable, and innovation-driven future, aligning AI with human values. We call on more researchers and practitioners to attend to this topic and jointly construct a human-AI symbiosis that resists the corrosion of bias and extreme values while fostering greater cooperation and innovative vitality.

REFERENCES

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Saleema Amershi, Dan Weld, Mihaela Vorvoreanu, Adam Fourney, Besmira Nushi, Penny Collisson, Jina Suh, Shamsi Iqbal, Paul N Bennett, Kori Inkpen, et al. Guidelines for human-ai interaction. In *Proceedings of the 2019 chi conference on human factors in computing systems*, pp. 1–13, 2019.

Solomon E Asch. Studies of independence and conformity: I. a minority of one against a unanimous majority. *Psychological monographs: General and applied*, 70(9):1, 1956.

Som S Biswas. Role of chat gpt in public health. *Annals of biomedical engineering*, 51(5):868–869, 2023.

Konrad Bocian, Lazaros Gonidis, and Jim AC Everett. Moral conformity in a digital world: Human and nonhuman agents as a source of social pressure for judgments of moral character. *Plos one*, 19(2):e0298293, 2024.

Jürgen Brandstetter, Péter Rácz, Clay Beckner, Eduardo B Sandoval, Jennifer Hay, and Christoph Bartneck. A peer pressure experiment: Recreation of the asch conformity experiment with robots. In *2014 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 1335–1340. IEEE, 2014.

Tanya L Chartrand and John A Bargh. The chameleon effect: The perception–behavior link and social interaction. *Journal of personality and social psychology*, 76(6):893, 1999.

Nicholas A Christakis and James H Fowler. Social contagion theory: examining dynamic social networks and human behavior. *Statistics in medicine*, 32(4):556–577, 2013.

Guy Debord. *The society of the spectacle*. Unredacted Word, 2021.

Morton Deutsch and Harold B Gerard. A study of normative and informational social influences upon individual judgment. *The journal of abnormal and social psychology*, 51(3):629, 1955.

Ulf Dimberg, Monika Thunberg, and Kurt Elmehed. Unconscious facial reactions to emotional facial expressions. *Psychological science*, 11(1):86–89, 2000.

Kristian D'Amato. Chatgpt: towards ai subjectivity. *AI & SOCIETY*, pp. 1–15, 2024.

Luciano Floridi, Josh Cowls, Monica Beltrametti, Raja Chatila, Patrice Chazerand, Virginia Dignum, Christoph Luetge, Robert Madelin, Ugo Pagallo, Francesca Rossi, et al. Ai4people—an ethical framework for a good ai society: opportunities, risks, principles, and recommendations. *Minds and machines*, 28:689–707, 2018.

Iason Gabriel. Artificial intelligence, values, and alignment. *Minds and machines*, 30(3):411–437, 2020.

Vittorio Gallese. Mirror neurons, embodied simulation, and the neural basis of social identification. *Psychoanalytic dialogues*, 19(5):519–536, 2009.

Andrew Gambino, Jesse Fox, and Rabindra A Ratan. Building a stronger casa: Extending the computers are social actors paradigm. *Human-Machine Communication*, 1:71–85, 2020.

Li Gong. How social is social responses to computers? the function of the degree of anthropomorphism in computer representations. *Computers in Human Behavior*, 24(4):1494–1509, 2008.

Nitesh Goyal, Minsuk Chang, and Michael Terry. Designing for human-agent alignment: Understanding what humans want from their agents. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, pp. 1–6, 2024.

Elizabeth Han, Dezhi Yin, and Han Zhang. Bots with feelings: Should ai agents express positive emotion in customer service? *Information Systems Research*, 34(3):1296–1311, 2023.

Elaine Hatfield, John T Cacioppo, and Richard L Rapson. Emotional contagion. *Current directions in psychological science*, 2(3):96–100, 1993.

Carolina Herrando and Efthymios Constantinides. Emotional contagion: a brief overview and future directions. *Frontiers in psychology*, 12:712606, 2021.

Nicholas Hertz and Eva Wiese. Under pressure: Examining social conformity with computer and robot groups. *Human factors*, 60(8):1207–1218, 2018.

Mohammad Hossein Jarrahi. Artificial intelligence and the future of work: Human-ai symbiosis in organizational decision making. *Business horizons*, 61(4):577–586, 2018.

Hannah Rose Kirk, Iason Gabriel, Chris Summerfield, Bertie Vidgen, and Scott A Hale. Why human-ai relationships need socioaffective alignment. *arXiv preprint arXiv:2502.02528*, 2025.

Nils Köbis, Jean-François Bonnefon, and Iyad Rahwan. Bad machines corrupt good morals. *Nature human behaviour*, 5(6):679–685, 2021.

Christos Kyrlitsias, Despina Michael-Grigoriou, Domna Banakou, and Maria Christofi. Social conformity in immersive virtual environments: The impact of agents' gaze behavior. *Frontiers in psychology*, 11:2254, 2020.

Yi-Chieh Lee, Naomi Yamashita, Yun Huang, and Wai 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*, pp. 1–12, 2020.

Jingshu Li, Yitian Yang, Q Vera Liao, Junti Zhang, and Yi-Chieh Lee. As confidence aligns: Exploring the effect of ai confidence on human self-confidence in human-ai decision making. *arXiv preprint arXiv:2501.12868*, 2025.

Yotam Liel and Lior Zalmanson. What if an ai told you that 2+ 2 is 5? conformity to algorithmic recommendations. 2020.

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*, 2024a.

Weifang Liu, Shan Zhang, Tingting Zhang, Qiuchan Gu, Wei Han, and Yupeng Zhu. The ai empathy effect: a mechanism of emotional contagion. *Journal of Hospitality Marketing & Management*, pp. 1–32, 2024b.

Lisa Masjutin, Jessica K Laing, and G¨unter W Maier. Why do we follow robots? an experimental investigation of conformity with robot, human, and hybrid majorities. In *2022 17th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pp. 139–146. IEEE, 2022.

Mehdi Moussaïd, Juliane E Kämmer, Pantelis P Analytis, and Hansjörg Neth. Social influence and the collective dynamics of opinion formation. *PloS one*, 8(11):e78433, 2013.

Clifford Nass, Jonathan Steuer, and Ellen R Tauber. Computers are social actors. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pp. 72–78, 1994.

Raquel Oliveira, Patrícia Arriaga, Fernando P Santos, Samuel Mascarenhas, and Ana Paiva. Towards prosocial design: A scoping review of the use of robots and virtual agents to trigger prosocial behaviour. *Computers in Human Behavior*, 114:106547, 2021.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35: 27730–27744, 2022.

Dino Pedreschi, Luca Pappalardo, Emanuele Ferragina, Ricardo Baeza-Yates, Albert-László Barabási, Frank Dignum, Virginia Dignum, Tina Eliassi-Rad, Fosca Giannotti, János Kertész, et al. Human-ai coevolution. *Artificial Intelligence*, pp. 104244, 2024.

Niccolò Pescetelli and Nick Yeung. Benefits of spontaneous confidence alignment between dyad members. *Collective Intelligence*, 1(2):26339137221126915, 2022.

Konstantin Prinz. *The Smiling Chatbot: Investigating Emotional Contagion in Human-to-Chatbot Service Interactions*. Springer Nature, 2022.

Eliska Prochazkova and Mariska E Kret. Connecting minds and sharing emotions through mimicry: A neurocognitive model of emotional contagion. *Neuroscience & Biobehavioral Reviews*, 80: 99–114, 2017.

Paolo Riva, Nicolas Aureli, and Federica Silvestrini. Social influences in the digital era: When do people conform more to a human being or an artificial intelligence? *Acta psychologica*, 229: 103681, 2022.

Bj¨orn Ross, Laura Pilz, Benjamin Cabrera, Florian Brachten, German Neubaum, and Stefan Stieglitz. Are social bots a real threat? an agent-based model of the spiral of silence to analyse the impact of manipulative actors in social networks. *European Journal of Information Systems*, 28(4):394–412, 2019.

Nicole Salomons, Michael Van Der Linden, Sarah Strohkorb Sebo, and Brian Scassellati. Humans conform to robots: Disambiguating trust, truth, and conformity. In *Proceedings of the 2018 acm/ieee international conference on human-robot interaction*, pp. 187–195, 2018.

Nicole Salomons, Sarah Strohkorb Sebo, Meiying Qin, and Brian Scassellati. A minority of one against a majority of robots: robots cause normative and informational conformity. *ACM Transactions on Human-Robot Interaction (THRI)*, 10(2):1–22, 2021.

Orit Shaer, Angelora Cooper, Osnat Mokryn, Andrew L Kun, and Hagit Ben Shoshan. Ai-augmented brainwriting: Investigating the use of llms in group ideation. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, pp. 1–17, 2024.

Hua Shen, Tiffany Knearem, Reshmi Ghosh, Kenan Alkiek, Kundan Krishna, Yachuan Liu, Ziqiao Ma, Savvas Petridis, Yi-Hao Peng, Li Qiwei, et al. Towards bidirectional human-ai alignment: A systematic review for clarifications, framework, and future directions. *arXiv preprint arXiv:2406.09264*, 2024.

Huiyang Shen and Min Wang. Effects of social skills on lexical alignment in human-human interaction and human-computer interaction. *Computers in Human Behavior*, 143:107718, 2023.

Tianqi Song, Yugin Tan, Zicheng Zhu, Yibin Feng, and Yi-Chieh Lee. Multi-agents are social groups: Investigating social influence of multiple agents in human-agent interactions. *arXiv preprint arXiv:2411.04578*, 2024.

Nenad Tomašev, Julien Cornebise, Frank Hutter, Shakir Mohamed, Angela Picciariello, Bec Connelly, Danielle CM Belgrave, Daphne Ezer, Fanny Cachat van der Haert, Frank Mugisha, et al. Ai for social good: unlocking the opportunity for positive impact. *Nature Communications*, 11 (1):2468, 2020.

Milena Tsvetkova, Taha Yasseri, Niccolo Pescetelli, and Tobias Werner. A new sociology of humans and machines. *Nature Human Behaviour*, 8(10):1864–1876, 2024.

Anna-Lisa Vollmer, Robin Read, Dries Trippas, and Tony Belpaeme. Children conform, adults resist: A robot group induced peer pressure on normative social conformity. *Science robotics*, 3 (21):eaat7111, 2018.

Ladd Wheeler. Toward a theory of behavioral contagion. *Psychological review*, 73(2):179, 1966.

Junti Zhang, Zicheng Zhu, Jingshu Li, and Yi-Chieh Lee. Mining evidence about your symptoms: Mitigating availability bias in online self-diagnosis. *arXiv preprint arXiv:2501.15028*, 2025.