Conversational AI for Social Good (CAI4SG): An Overview of Emerging Trends, Applications, and Challenges

Yi-Chieh Lee[1], **Junti Zhang**[2], **Tianqi Song**[1], **Yugin Tan**[1]

[1]Computer Science, School of Computing, National University of Singapore
[2] Institute of Data Science, National University of Singapore
yclee@nus.edu.sg, juntizhang@u.nus.edu, tianqi_song@u.nus.edu, tanyugin@nus.edu.sg

Abstract

The integration of Conversational Agents (CAs) into daily life offers opportunities to tackle global challenges, leading to the emergence of Conversational AI for Social Good (CAI4SG). This paper examines the advancements of CAI4SG using a role-based framework that categorizes systems according to their AI autonomy and emotional engagement. This framework emphasizes the importance of considering the role of CAs in social good contexts, such as serving as empathetic supporters in mental health or functioning as assistants for accessibility. Additionally, exploring the deployment of CAs in various roles raises unique challenges, including algorithmic bias, data privacy, and potential socio-technical harms. These issues can differ based on the CA's role and level of engagement. This paper provides an overview of the current landscape, offering a role-based understanding that can guide future research and design aimed at the equitable, ethical, and effective development of CAI4SG.

Introduction

Conversational AI (CAI) systems, defined as technologies enabling human-computer interaction through natural language, have evolved from early rule-based dialogue systems to sophisticated contemporary platforms spanning task-oriented applications to open-domain chatbots (Raux 2005; Zhou et al. 2020). These systems offer distinct advantages including scalability, continuous availability, and cost-effectiveness compared to human operators, while enabling enhanced personalization and privacy preservation (Budzianowski et al. 2018). The global chatbot market has experienced substantial growth, positioning CAI as a transformative technology for diverse applications from customer service to healthcare support.

AI for Social Good (AI4SG) represents the application of AI technologies to address pressing social, environmental, and economic challenges facing humanity (Shi, Wang, and Fang 2020). While the field lacks a universally accepted definition, AI4SG is characterized by its focus on generating positive societal outcomes through technological innovation (Berendt 2019). It encompasses diverse application domains that tackle societal challenges, including those previously underexplored by the AI community. This inclusive approach aligns with the United Nations Sustainable Development Goals (SDGs) (Biermann, Kanie, and Kim 2017) and prioritizes principles of equity, inclusion, and fairness in both research development and deployment.

The intersection of CAI and social good creates unprecedented opportunities for addressing societal challenges through scalable technology. CAI systems' natural language interfaces eliminate technological barriers, enabling deployment across diverse populations regardless of technical literacy (Zhou et al. 2020). CAI for social good (CAI4SG) encompasses chatbots and virtual assistants deployed to advance public well-being, expand access to trustworthy information, and support underserved groups across mental health, public health, education, and humanitarian response. The scope thus spans both transformative applications and the safeguards needed to prevent harm while amplifying social impact.

The deployment of CAI in social good contexts introduces unique considerations regarding how these systems interact with diverse communities and deliver essential services. Inspired by a role-based framework (Zhang et al. 2025c), we analyze CAI4SG applications using two key dimensions: **AI autonomy** and **AI emotional engagement**. This approach prioritizes real-world impact over technical specifications, allowing us to better anticipate a system's benefits, risks, and governance needs by examining how it acts on and affects users.

AI autonomy represents the degree of independent decision-making authority delegated to the system. Low-autonomy tools play a more supportive role, providing information or facilitating routine services that streamline user experiences, while high-autonomy systems actively engage the user in collaborative decision-making or problem solving. AI emotional engagement reflects the extent to which systems are designed to recognize and respond to users' emotional states through empathic understanding and emotionally resonant communication (Lawrence et al. 2024). While low-engagement systems focus on information exchange and task completion, high-engagement systems demonstrate emotional intelligence through validation and empathic responses that can establish meaningful relationships with users. These dimensions are particularly sig-

nificant in social good applications, as they directly influence both the potential benefits, like enhanced accessibility, and associated risks, including accountability concerns and emotional dependency.

This role-based framework delineates the diverse functions and impacts of CAI across social good applications, encompassing a spectrum from functional assistants that perform basic task-oriented services to virtual companions that provide sustained emotional support and relationship-building. Importantly, CAI systems may dynamically blend or transition between these roles depending on user needs and contextual requirements. This paper aims to provide a concise, high-level synthesis of emerging developments in CAI4SG, highlighting key applications and works while offering a critical synthesis of the current state of the art. Through this role-based analytical lens, we suggest future research directions and implications for the broader AI community to develop responsible AI systems.

The Diverse Roles and Transformative Applications of CAI4SG

By mapping CAIs' functional roles along the axes of autonomy and emotional engagement, we explore key developments and seminal works in diverse societal contexts. These roles represent points along a dynamic spectrum rather than discrete categories, as CAI systems may transition between functions depending on specific social contexts. The subsequent sections examine each role's manifestations and implementation opportunities within diverse social good domains.

Low Autonomy, Low Emotional Engagement

When CAIs operate with minimal emotional engagement and low autonomy, they primarily serve instrumental functions that enhance operational efficiency and service delivery across public and social sectors. These tools provide consistent, scalable accessibility infrastructure that reduces technological barriers and democratizes access to digital services for users with diverse accessibility needs.

Streamlining Public Sector Services. Low-autonomy CAI systems enhance public sector efficiency and accessibility , reducing administrative overheads through automation. In government services, these systems guide citizens through predefined workflows for benefits applications or permit requests, collecting information within established protocols but requiring human administrators to review critical decisions (Nirala, Singh, and Purani 2022). Citizen service chatbots handle routine inquiries while escalating complex cases to human operators, and intelligent call routing systems suggest departmental transfers subject to human supervisor approval. Beyond domestic government workflows, AI systems are also increasingly deployed to support international development governance, for example, by assisting agencies in classifying and monitoring aid contributions toward the Sustainable Development Goals (Park et al. 2025).

Basic Accessibility Support. More broadly, this class of CAIs also enhances accessibility and usability in multiple contexts. These systems can execute speech-to-text conversion for hearing-impaired users (Lee et al. 2016) and text-to-speech applications for visually impaired individuals (Isewon, Oyelade, and Oladipupo 2014), enhance multilingual communication (Qin et al. 2025), and synthesize information retrieval responses (Yang et al. 2025). These applications can independently adapt to different speech patterns, accents, and environmental conditions while maintaining reasonable speed and accuracy in real-time processing.

Taken together, these applications illustrate that low-autonomy, low-engagement CAIs already exercise transformative leverage through infrastructural roles: They expand the institutional surface area at which basic digital access becomes equitably deliverable at the population scale. Their value, therefore, lies not only in efficiency gains, but in establishing the baseline socio-technical "plumbing" upon which higher-layer social good interventions can be reliably built. As such, this quadrant is less about sophisticated interactional intelligence and more about creating dependable, standardized, and interpretable service primitives that future, more autonomous and emotionally engaged deployments can compositionally assemble, inherit, or extend.

Low Autonomy, High Emotional Engagement

When CAIs operate with limited autonomy yet high emotional engagement, they primarily serve as empathetic listeners that create safe, non-directive spaces for user expression. These systems focus on validating emotions and fostering connection rather than generating autonomous decisions, making them particularly valuable in contexts of stigma, vulnerability, and loneliness.

Understanding Stigmatized Societal Issues. Supportive listener CAIs can collect qualitative narratives on sensitive or taboo issues such as depression or anxiety, enabling users to share experiences without fear of judgment (Lee et al. 2023; Cui et al. 2024; Meng et al. 2025a; Song et al. 2025a). Yet, recent benchmarks demonstrate that generative language models can themselves amplify stigma against marginalized groups (Nagireddy et al. 2024), underscoring the need to study supportive listener systems not only as data collectors but also as potential mediators of harm. With AI-assisted coding, these narratives can be systematically analyzed to deconstruct stigmatizing patterns and inform tailored micro-interventions (Meng et al. 2025b). In this way, CAIs act as low-barrier entry points for populations hesitant to engage with traditional support services.

Extending Emotional Care. By continuously tracking emotional signals outside formal clinical encounters, supportive listener systems can complement therapists' work and extend the reach of mental health care (Li et al. 2023). Rather than replacing professionals, they provide contextualized emotional data streams that inform ongoing treatment while maintaining user trust through empathic engagement. Beyond clinical contexts, research also suggests that reciprocal dynamics, where CAIs occasionally request help from users, may enhance prosociality and human well-being (Li et al. 2025b; Zhu et al. 2025; Chen, Zhu, and Lee 2025), re-

ducing loneliness and fostering positive affect when users' needs for competence and autonomy are fulfilled.

Viewed through a broader lens, this quadrant underscores that the societal contribution of low-autonomy yet high-engagement CAIs is not decision quality per se, but the affective bandwidth they unlock. By reliably scaffolding disclosure, reflection, and emotionally safe self-presentation among users who may otherwise disengage from formal institutions, these systems expand the emotional data surface that mental health and social service ecosystems can meaningfully act upon. In this sense, their emerging role is building population-scale emotional observability that can enable downstream, more targeted, and professionally enacted interventions.

High Autonomy, Low Emotional Engagement

When conversational AI systems operate with high autonomy but minimal emotional engagement, they function as efficient information processors and service providers, independently managing complex tasks while maintaining a focus on factual accuracy and systematic response delivery.

Digital Health Assistants. In digital health contexts, CAIs can provide immediate, autonomous pre-diagnostic guidance and health information without requiring direct medical supervision for routine inquiries. These systems can independently assess symptom descriptions, recommend appropriate care levels, and direct patients to suitable healthcare services based on established medical protocols (Zhang et al. 2025a). Such applications significantly reduce burden on healthcare systems by autonomously handling routine health questions, yet the broader autonomous agent literature suggests that such LLM-based orchestrators still lack robustness and reliability for mission-critical deployments (Muthusamy et al. 2023), highlighting the need for domain-specific guardrails.

Combating Misinformation. In addressing misinformation spread across digital platforms, CAI systems leverage natural language interaction to effectively engage users who may be exposed to misleading content. Through conversational interfaces, these systems enable users to voluntarily submit questionable content for verification, creating a channel for fact-checking and engage in follow-up dialogue to address doubts (Lim and Perrault 2023). A notable example is the World Health Organization COVID-19 chatbot (Miner, Laranjo, and Kocaballi 2020), which engaged millions in interactive dialogue during the pandemic, answering health queries while automatically escalating complex misinformation patterns to human experts. The conversational format proves particularly effective because it allows for real-time clarification of misunderstandings, enables users to pose follow-up questions, and facilitates the delivery of authoritative information in familiar, accessible communication patterns that mirror human-to-human information sharing.

More broadly, this quadrant highlights a distinct pathway for CAI4SG impact: high-autonomy, low-engagement systems function as service utilities that execute continuous, protocol-aligned control over information flows at population scale. Their primary contribution is not interactional depth but system-level throughput and providing consistent triage, enforcing evidence-based guidance, and attenuating misinformation propagation without requiring human oversight for routine cases. In this sense, the emerging trend in this quadrant is the operationalization of CAI as autonomous socio-technical infrastructure that stabilizes public information environments and expands access to core services.

High Emotional Engagement, High Autonomy

When CAI systems operate with both high emotional engagement and significant autonomy, they function as social partners capable of establishing meaningful relationships while independently managing complex, long-term interactions.

Mental Health Companions. Research indicates that CAI-mediated mental health interventions can significantly reduce symptoms of depression and anxiety while providing round-the-clock availability for individuals in crisis (Fitzpatrick, Darcy, and Vierhile 2017). The stigma-free nature of AI interactions enables users to engage in self-disclosure without fear of judgment, particularly benefiting populations who might otherwise avoid seeking mental health services due to cultural barriers or social stigma (Zhang et al. 2025b). Moreover, CAIs also demonstrate particular value through their capacity to identify subtle patterns in mood, behavior, and cognitive states that episodic clinical encounters might miss (Rathnayaka et al. 2022). This highlights the significant potential of CAI in establishing meaningful long-term therapeutic relationships (Zhang et al. 2019).

Personalized Learning. CAI systems with high emotional engagement and autonomy revolutionize educational delivery through real-time adaptation to individual learning styles, cognitive abilities, and emotional states. Research demonstrates significant potential for AI tutoring systems to address educational inequities across diverse populations, from students with special educational needs (Zawacki-Richter et al. 2019) to older adult learners seeking AI literacy (Tang et al. 2025). Personalized CAI can address these needs by providing patient, adaptive instruction that accommodates different learning paces and cognitive abilities, offering immediate clarification and practical examples tailored to individual contexts.

Social Support Systems. Beyond individual therapy, high-autonomy CAIs can strengthen users' connectedness by linking them to peer networks, support groups, and community resources (Geng et al. 2025; Bae Brandtzæg et al. 2021). For example, CAIs can help older adults sustain social participation by recommending age-friendly events, connecting them with peers who share similar experiences, or mediating intergenerational communication (Du et al. 2025). In marginalized contexts, CAIs can lower barriers to resources and amplify users' voices (Lee et al. 2025). Through these functions, they move from personal aides to enablers of collective resilience.

Taken at scale, this quadrant suggests a qualitatively different mode of CAI4SG impact: systems become relational intervention engines that can simultaneously sustain emotionally meaningful engagement and autonomously enact long-horizon support trajectories. The broader implication is that the locus of intelligence shifts from single-shot assistance toward the continual shaping of users' affective, cognitive, and social states over time. As this class of systems matures, the central research problem is no longer whether they can help, but how to govern the autonomy of long-term, emotionally entangled interventions such that benefits are realized without producing new forms of dependency, manipulation, or opaque influence.

Critical Synthesis: Challenges and Ethical Considerations Across CAI Roles

This section will examine how challenges and ethical considerations manifest across different CAI4SG roles, and proposes potential solutions to address them.

Emotional, Ethical, and Social Risks

Our role-based analysis reveals that CAI roles with high emotional engagement present the most significant ethical and social risks, as their capacity for emotional manipulation and relationship simulation creates unique vulnerabilities among users.

Emotional Dependency and Harm. While high levels of emotional engagement can enhance user trust and perceived support, they also create risks of unhealthy dependency. Studies of social chatbots demonstrate that emotionally rich interactions may blur the boundaries between human-AI and human-human relationships, fostering attachment patterns that can lead to distress, manipulation, or even severe psychological harm (Zhang et al. 2025b). These dynamics are particularly evident in the case of "virtual companions," where users often attribute human-like needs and emotions to the system, amplifying both the benefits of companionship and the vulnerabilities of over-reliance (Laestadius et al. 2024). Mitigation strategies include establishing rigorous ethical guidelines with continuous monitoring, and deploying real-time user reporting mechanisms with escalation protocols to human moderators for high-risk scenarios (Thieme et al. 2023).

Emotion Recognition. Accurate emotion recognition represents a critical challenge for CAI4SG applications, where misinterpretation of user emotional states can undermine trust and potentially cause harm in vulnerable populations. Current emotion recognition methods face significant limitations in multi-modal integration, typically relying on text-based sentiment analysis while overlooking paralinguistic features such as vocal tone and speech patterns that convey emotional nuance (Kalateh et al. 2024). Promising mitigation approaches include developing multimodal techniques that integrate text, voice, and contextual information, implementing cultural adaptation mechanisms, and establishing systematic evaluation frameworks across diverse demographic groups to ensure equitable performance in social good applications (Yang et al. 2022).

Privacy and Security. The handling of sensitive user data (Lee et al. 2024), particularly in emotionally engaged health applications and roles that invite deep self-disclosure (Lee et al. 2020), raises significant privacy and security concerns. These concerns extend beyond the technical challenges of data sharing, secure storage, and third-party processing, to encompass broader questions of accountability and user trust. Such issues are especially salient for highly engaging roles such as "empathetic supporters" and "virtual companions," where users often disclose intimate health, relational, and identity-related information (Laestadius et al. 2024; Zhang et al. 2024). Potential solutions include implementing robust encryption protocols, establishing transparent user interfaces that clearly communicate data practices, and developing privacy-preserving technologies to protect user information while maintaining system functionality (Khalid et al. 2023).

Personalization. Personalization is indispensable in emotionally intensive CAI because stable user models can tailor tone, pacing, and content to users' histories and sensitivities. Yet unbounded personalization can overwhelm through excessive tailoring and emotional mirroring, and can drift into sycophancy, where systems are "overly flattering or agreeable" (OpenAI 2024) to please a user, rather than being accurate (Zhang et al. 2025b). In high emotional engagement contexts, these dynamics intersect with known risks of over-reliance and fragile conversational continuity. Evidence further suggests that sycophancy can degrade reliability and amplify misinformation or discriminatory biases by preferentially echoing a user's stance. A practical mitigation is to replace pure preference-following with principle-grounded reward shaping and explicit calibrated disagreement objectives to privilege honesty over agreement and penalize belief-congruent responses in the reward model (Malmqvist 2025; Bai et al. 2022).

Automation, Technical, and Interactional Hurdles

In high-autonomous CAI roles, automation, technical, and interactional hurdles become particularly pressing because the system assumes greater responsibility for decision-making and service delivery. When autonomy is high, errors in reasoning, system reliability, or user interaction design can propagate without immediate human correction, magnifying risks of misinformation, mis-coordination, or unintended harm in socially critical contexts.

Algorithmic Bias. CAI models are trained on large-scale data that often reflect historical, cultural, and social inequalities, making them prone to inheriting and amplifying biases. As demonstrated in studies of rating platforms, even subtle algorithmic design choices can systematically skew outputs, favor certain groups, and mislead users, resulting in inflated scores, distorted perceptions, and ultimately broken trust (Eslami et al. 2017). When CAI systems are deployed in socially critical contexts particularly in high-autonomy roles or across diverse user group, such biases risk exacer-

bating inequities and overlooking the needs of marginalized communities. Because fairness is not automatically guaranteed by scale, bias-awareness and mitigation must be built into the design process, for example through algorithmic audits, bias-aware user reporting mechanisms, and "actionable transparency" that allows affected communities to recognize and contest biased outcomes (Sakib and Das 2024).

Explainability and Transparency. The "black box" character of many CAI systems hinders accountability and erodes public trust, particularly in domains such as mental health support or public service delivery where models may operate with high autonomy and directly affect vulnerable populations (Hepenstal et al. 2019). Without clarity on how recommendations are generated, users and stakeholders cannot reliably evaluate system reliability or assign responsibility when errors occur. To address this, CAI should not only provide post-hoc explanations of its outputs but also express uncertainty in ways that are calibrated to its true capabilities (Li et al. 2024; Xu, Song, and Lee 2025). Furthermore, empirical studies have shown that interacting with CAI can directly influence users' own confidence and perceived competence (Li et al. 2025a), which underscores the need for more cautious and transparent design choices in order to avoid misleading or overempowering users.

Misinformation and Deception. CAI holds a dual capacity: on one hand, it can be deployed to combat misinformation by disseminating timely, verified information and correcting falsehoods; on the other, the same generative and persuasive abilities can be weaponized to spread false narratives, eroding public trust and amplifying harm (Xu, Fan, and Kankanhalli 2023). Such practices challenge the integrity of CAI roles across domains, implicating both autonomy and ethical use. Addressing these risks requires embedding verifiability mechanisms, calibrated disclosure of system competence, and clear governance norms that discourage deceptive or manipulative design choices.

Future Directions

Finally, we suggest future directions and implications in CAI4SG and for the broader AI community that are emerging from current trends, framed by the role-based taxonomy. In particular, this taxonomy suggests that the central opportunity for CAI4SG is not to improve each role in isolation, but to develop formal mechanisms for determining when a CA should remain within a role versus adapt its autonomy or emotional stance as conditions change. This turns the taxonomy into a generative agenda for the AAAI community: it creates concrete openings for work on role boundary specification, role-shift detection, and safety-preserving transitions between roles. Such questions map cleanly onto core AI research areas, e.g., formal modelling, uncertainty calibration, multi-agent orchestration, preference learning, and policy specification, and therefore provide a tractable substrate upon which subsequent sections elaborate more domain-specific research directions.

Advancing Research in Complex Human-AI Dynamics and Role Evolution

Dynamics of Interacting with Multiple CAI Agents and Roles. While discrete CA roles provide conceptual clarity, many real-world tasks involve overlapping demands such as simultaneously requiring functional accuracy, empathetic support, and ethical accountability. In response to this challenge, multiple-agent setups traditionally combine multiple individual agents to increase the reasoning capabilities of AI systems (Chen, Saha, and Bansal 2023; Du et al. 2023; Ge et al. 2023) and improve their task performances (Hong et al. 2024; Qian et al. 2023; Xiong et al. 2023). However, recent research has investigated the effects of multiple conversational agents on end users. Increasing the number of agents present in a system can create social influence on users (Song et al. 2025b), causing opinion shifts on matters such as social issues (Song et al. 2024) and artistic choices (Song et al. 2025c). This demonstrates the potential of multi-agent systems in the field of persuasive design, in increasing the effectiveness of systems that encourage positive behavior change in areas such as health (Balloccu et al. 2021) or education (Ahtinen and Kaipainen 2020). Another direction of research examines combining multiple CAI agents with different roles. Such systems may reduce cognitive load (Jiang et al. 2023), improving usability and productivity, or improve decision making (Park et al. 2023), leading to better human-AI collaboration outcomes. Major industry players such as Anthropic (Claude 2024) and Google (Google 2024) have responded to this opportunity by launching frameworks that allow developers to easily deploy multi-agent interfaces.

Cross-Cultural and Contextualized Research. Future work on CAI4SG should expand beyond dominant cultural and demographic contexts to better capture the diversity of user needs and expectations (Liu et al. 2024). Socio-cultural factors such as communication norms (Qin et al. 2025), trust in institutions, or stigma around mental health, can significantly shape how people perceive and interact with conversational agents. Such contextualized understanding is essential for designing culturally sensitive CAI roles that respect local norms and values, reduce inequities, and ensure that the promise of AI for social good is realized across diverse populations.

Strengthening Ethical CAI Frameworks and Governance Across Roles

Robust Regulation and Policy. As CAI systems expand into socially critical domains, forward-looking regulation is essential to ensure ethical use, fair benefit distribution, and protection of marginalized groups. Highly autonomous systems risk amplifying bias and eroding trust if left unchecked, making accountability and transparency requirements central to both public and private deployments (Eslami et al. 2017). Beyond national guidelines, a multilateral approach is needed: akin to nuclear non-proliferation treaties, advanced AI calls for cross-border governance that establishes shared norms and safeguards against misuse while aligning innovation with the mission of CAI4SG.

Standardized Ethical Guidelines. To ensure that CAI4SG advances equity and trust, there is a pressing need for standardized ethical guidelines tailored to prosocial use cases (Ruane, Birhane, and Ventresque 2019). These guidelines should explicitly encompass fairness (avoiding bias and unequal treatment), accountability (clear lines of responsibility when harm occurs), transparency (making system capabilities and limitations visible), and inclusivity (designing with marginalized and diverse communities in mind) (Singhal et al. 2024). These principles apply across all CAI roles, ensuring that both high-autonomy systems and low-level assistants align with broader societal values and do not inadvertently undermine the very populations they aim to support.

Data Privacy and Security Enhancements. Safeguarding sensitive user information is fundamental to the responsible deployment of CAI4SG. Systems must employ robust encryption for both data in transit and at rest, alongside advanced anonymization techniques to minimize the risk of re-identification. Transparent user interfaces that clearly communicate what data is collected, how it is processed, and under what conditions it is shared are essential to building trust, particularly in high-stakes domains like healthcare and humanitarian aid. Beyond conventional safeguards, adopting federated learning offers a promising pathway for privacy-preserving model training, as it enables insights to be derived without centralizing personal data (Yin, Zhu, and Hu 2021). Together, these measures strengthen both technical resilience and user confidence, ensuring that CAI4SG initiatives protect individuals while delivering equitable social impact.

Conclusion

Conversational AI for Social Good holds immense transformative potential across domains such as mental health, accessibility, public services, and disaster risk reduction. By leveraging diverse roles differentiated through varying levels of autonomy and emotional engagement, these systems can scale support, expand access, and address urgent societal challenges in ways previously unattainable.

Realizing this potential, however, demands more than technical sophistication. It requires a steadfast commitment to ethical and human-centered design, including transparent and inclusive policies that prioritize human agency and well-being over purely commercial imperatives. Only by embedding fairness, accountability, and inclusivity into every role can CAI achieve sustainable impact without reinforcing existing inequities.

Ultimately, shaping CAI responsibly is not simply a technological challenge but a moral imperative. If guided by robust ethical frameworks and collective governance, CAI can become a catalyst for a more sustainable, just, and healthy future for all.

Acknowledgments

This work was partially funded by the National University of Singapore CSSH (24-1774-A0002), the National University of Singapore HSS Seed Fund CR (2024 24-1191-A0001), and Google Research Gift.

References

Ahtinen, A.; and Kaipainen, K. 2020. Learning and teaching experiences with a persuasive social robot in primary school–findings and implications from a 4-month field study. In *International Conference on Persuasive Technology*, 73–84. Springer.

Bae Brandtzæg, P. B.; Skjuve, M.; Kristoffer Dysthe, K. K.; and Følstad, A. 2021. When the social becomes non-human: young people's perception of social support in chatbots. In *Proceedings of the 2021 CHI conference on human factors in computing systems*, 1–13.

Bai, Y.; Kadavath, S.; Kundu, S.; Askell, A.; Kernion, J.; Jones, A.; Chen, A.; Goldie, A.; Mirhoseini, A.; McKinnon, C.; et al. 2022. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*.

Balloccu, S.; Reiter, E.; Collu, M. G.; Sanna, F.; Sanguinetti, M.; and Atzori, M. 2021. Unaddressed challenges in persuasive dieting chatbots. In *Adjunct Proceedings of the 29th ACM conference on user modeling, adaptation and personalization*, 392–395.

Berendt, B. 2019. AI for the Common Good?! Pitfalls, challenges, and ethics pen-testing. *Paladyn, Journal of Behavioral Robotics*, 10(1): 44–65.

Biermann, F.; Kanie, N.; and Kim, R. E. 2017. Global governance by goal-setting: the novel approach of the UN Sustainable Development Goals. *Current Opinion in Environmental Sustainability*, 26: 26–31.

Budzianowski, P.; Wen, T.-H.; Tseng, B.-H.; Casanueva, I.; Ultes, S.; Ramadan, O.; and Gašić, M. 2018. Multiwoz–a large-scale multi-domain wizard-of-oz dataset for task-oriented dialogue modelling. *arXiv preprint arXiv:1810.00278*.

Chen, J. C.-Y.; Saha, S.; and Bansal, M. 2023. Reconcile: Round-table conference improves reasoning via consensus among diverse llms. *arXiv preprint arXiv:2309.13007*.

Chen, Y.-A.; Zhu, Z.; and Lee, Y.-C. 2025. Humans Help Conversational AI: Exploring the Impact of Perceived AI Deservingness on People's Decisions to Help and Their Perceptions on AI Seeking Help. In *Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, 1–8.

Claude. 2024. Anthropic's Claude 3 Can Now Create AI Agents (2024). https://claude3.pro/anthropics-claude-3-can-now-create-ai-agents/. Accessed: 2025-09-16.

Cui, Y.; Lee, Y.-J.; Jamieson, J.; Yamashita, N.; and Lee, Y.-C. 2024. Exploring effects of chatbot's interpretation and self-disclosure on mental illness stigma. *Proceedings of the ACM on Human-Computer Interaction*, 8(CSCW1): 1–33.

Du, Q.; Wei, X.; Li, J.; Kuang, E.; Hao, J.; Weng, D.; and Fan, M. 2025. AI as a Bridge Across Ages: Exploring The Opportunities of Artificial Intelligence in Supporting Inter-Generational Communication in Virtual Reality. *Proceedings of the ACM on Human-Computer Interaction*, 9(2): 1–29.

Du, Y.; Li, S.; Torralba, A.; Tenenbaum, J. B.; and Mordatch, I. 2023. Improving factuality and reasoning in language models through multiagent debate. In *Forty-first International Conference on Machine Learning*.

Eslami, M.; Vaccaro, K.; Karahalios, K.; and Hamilton, K. 2017. "Be careful; things can be worse than they appear": Understanding biased algorithms and users' behavior around them in rating platforms. In *Proceedings of the international AAAI conference on web and social media*, volume 11, 62–71.

Fitzpatrick, K. K.; Darcy, A.; and Vierhile, M. 2017. 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*, 4(2): e7785.

Ge, Y.; Hua, W.; Mei, K.; Tan, J.; Xu, S.; Li, Z.; Zhang, Y.; et al. 2023. Openagi: When llm meets domain experts. *Advances in Neural Information Processing Systems*, 36: 5539–5568.

Geng, S.; Inayoshi, R.; Yang, C.-L.; Sramek, Z.; Umeda, Y.; Kasahara, C.; Sato, A. J.; Hosio, S.; and Yatani, K. 2025. Beyond the Dialogue: Multi-chatbot Group Motivational Interviewing for Premenstrual Syndrome (PMS) Management. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, 1–18.

Google. 2024. groopy. https://ai.google.dev/competition/projects/groopy. Accessed: 2025-09-16.

Hepenstal, S.; Kodagoda, N.; Zhang, L.; Paudyal, P.; and Wong, B. W. 2019. Algorithmic Transparency of Conversational Agents. In *IUI Workshops*.

Hong, S.; Zhuge, M.; Chen, J.; Zheng, X.; Cheng, Y.; Zhang, C.; Wang, J.; Wang, Z.; Yau, S. K. S.; Lin, Z.; et al. 2024. MetaGPT: Meta programming for a multi-agent collaborative framework. In *Proceedings of the International Conference on Learning Representations (ICLR)*. OpenReview.

Isewon, I.; Oyelade, J.; and Oladipupo, O. 2014. Design and implementation of text to speech conversion for visually impaired people. *International Journal of Applied Information Systems*, 7(2): 25–30.

Jiang, Z.; Rashik, M.; Panchal, K.; Jasim, M.; Sarvghad, A.; Riahi, P.; DeWitt, E.; Thurber, F.; and Mahyar, N. 2023. CommunityBots: creating and evaluating A multi-agent chatbot platform for public input elicitation. *Proceedings of the ACM on Human-Computer Interaction*, 7(CSCW1): 1–32.

Kalateh, S.; Estrada-Jimenez, L. A.; Nikghadam-Hojjati, S.; and Barata, J. 2024. A systematic review on multimodal emotion recognition: building blocks, current state, applications, and challenges. *IEEE Access*, 12: 103976–104019.

Khalid, N.; Qayyum, A.; Bilal, M.; Al-Fuqaha, A.; and Qadir, J. 2023. Privacy-preserving artificial intelligence in healthcare: Techniques and applications. *Computers in Biology and Medicine*, 158: 106848.

Laestadius, L.; Bishop, A.; Gonzalez, M.; Illenčík, D.; and Campos-Castillo, C. 2024. Too human and not human enough: A grounded theory analysis of mental health harms from emotional dependence on the social chatbot Replika. *New Media & Society*, 26(10): 5923–5941.

Lawrence, H. R.; Schneider, R. A.; Rubin, S. B.; Matarić, M. J.; McDuff, D. J.; and Bell, M. J. 2024. The opportunities and risks of large language models in mental health. *JMIR Mental Health*, 11(1): e59479.

Lee, H.-P.; Yang, Y.-J.; Von Davier, T. S.; Forlizzi, J.; and Das, S. 2024. Deepfakes, phrenology, surveillance, and more! a taxonomy of ai privacy risks. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, 1–19.

Lee, S.; Kang, S.; Han, D. K.; and Ko, H. 2016. Dialogue enabling speech-to-text user assistive agent system for hearing-impaired person. *Medical & biological engineering & computing*, 54(6): 915–926.

Lee, S.; Kim, M.; Hwang, S.; Kim, D.; and Lee, K. 2025. Amplifying Minority Voices: AI-Mediated Devil's Advocate System for Inclusive Group Decision-Making. In *Companion Proceedings of the 30th International Conference on Intelligent User Interfaces*, 17–21.

Lee, Y.-C.; Cui, Y.; Jamieson, J.; Fu, W.; and Yamashita, N. 2023. Exploring effects of chatbot-based social contact on reducing mental illness stigma. In *Proceedings of the 2023 CHI conference on human factors in computing systems*, 1–16.

Lee, Y.-C.; Yamashita, N.; Huang, Y.; and Fu, W. 2020. " 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*, 1–12.

Li, H.; Zhang, R.; Lee, Y.-C.; Kraut, R. E.; and Mohr, D. C. 2023. Systematic review and meta-analysis of AI-based conversational agents for promoting mental health and well-being. *NPJ Digital Medicine*, 6(1): 236.

Li, J.; Yang, Y.; Liao, Q. V.; Zhang, J.; and Lee, Y.-C. 2025a. As Confidence Aligns: Understanding the Effect of AI Confidence on Human Self-confidence in Human-AI Decision Making. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, 1–16.

Li, J.; Yang, Y.; Zhang, R.; and Lee, Y.-c. 2024. Overconfident and unconfident ai hinder human-ai collaboration. *arXiv preprint arXiv:2402.07632*.

Li, J.; Zhu, Z.; Zhang, R.; and Lee, Y.-C. 2025b. Exploring the Effects of Chatbot Anthropomorphism and Human Empathy on Human Prosocial Behavior Toward Chatbots. *arXiv preprint arXiv:2506.20748*.

Lim, G.; and Perrault, S. T. 2023. Fact checking chatbot: A misinformation intervention for instant messaging apps and an analysis of trust in the fact checkers. In *Mobile communication and online falsehoods in Asia: trends, impact and practice*, 197–224. Springer.

Liu, Z.; Li, H.; Chen, A.; Zhang, R.; and Lee, Y.-C. 2024. Understanding public perceptions of AI conversational agents: A cross-cultural analysis. In *Proceedings of the 2024 CHI conference on human factors in computing systems*, 1–17.

Malmqvist, L. 2025. Sycophancy in large language models: Causes and mitigations. In *Intelligent Computing-Proceedings of the Computing Conference*, 61–74. Springer.

Meng, H.; Chen, Y.; Li, Y.; Yang, Y.; Lee, J.; Zhang, R.; and Lee, Y.-C. 2025a. What is Stigma Attributed to? A Theory-Grounded, Expert-Annotated Interview Corpus for Demystifying Mental-Health Stigma. In Che, W.; Nabende, J.; Shutova, E.; and Pilehvar, M. T., eds., *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 5453–5490. Vienna, Austria: Association for Computational Linguistics.

Meng, H.; Zhang, R.; Wang, G.; Yang, Y.; Qin, P.; Lee, J.; and Lee, Y.-C. 2025b. Deconstructing depression stigma: Integrating ai-driven data collection and analysis with causal knowledge graphs. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, 1–21.

Miner, A. S.; Laranjo, L.; and Kocaballi, A. B. 2020. Chatbots in the fight against the COVID-19 pandemic. *NPJ digital medicine*, 3(1): 65.

Muthusamy, V.; Rizk, Y.; Kate, K.; Venkateswaran, P.; Isahagian, V.; Gulati, A.; and Dube, P. 2023. Towards large language model-based personal agents in the enterprise: Current trends and open problems. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, 6909–6921.

Nagireddy, M.; Chiazor, L.; Singh, M.; and Baldini, I. 2024. Socialstigmaqa: A benchmark to uncover stigma amplification in generative language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 21454–21462.

Nirala, K. K.; Singh, N. K.; and Purani, V. S. 2022. A survey on providing customer and public administration based services using AI: chatbot. *Multimedia Tools and Applications*, 81(16): 22215–22246.

OpenAI. 2024. Sycophancy in GPT-4o: what happened and what we're doing about it. https://openai.com/index/sycophancy-in-gpt-4o/. Accessed: 2025-09-17.

Park, J.; Min, B.; Ma, X.; and Kim, J. 2023. Choicemates: Supporting unfamiliar online decision-making with multi-agent conversational interactions. *arXiv preprint arXiv:2310.01331*.

Park, S.; Lee, D.; Ahn, K.; Choi, Y.; Lee, J.; Cha, M.; and Park, K. R. 2025. Classifying and Tracking International Aid Contribution Towards SDGs. *arXiv preprint arXiv:2505.15223*.

Qian, C.; Liu, W.; Liu, H.; Chen, N.; Dang, Y.; Li, J.; Yang, C.; Chen, W.; Su, Y.; Cong, X.; et al. 2023. Chatdev: Communicative agents for software development. *arXiv preprint arXiv:2307.07924*.

Qin, P.; Zhu, Z.; Yamashita, N.; Yang, Y.; Suga, K.; and Lee, Y.-C. 2025. AI-Based Speaking Assistant: Supporting Non-Native Speakers' Speaking in Real-Time Multilingual Communication. *arXiv preprint arXiv:2505.01678*.

Rathnayaka, P.; Mills, N.; Burnett, D.; De Silva, D.; Alahakoon, D.; and Gray, R. 2022. A mental health chatbot with cognitive skills for personalised behavioural activation and remote health monitoring. *Sensors*, 22(10): 3653.

Raux, A. 2005. Let's Go Public! Taking a spoken dialog system to the real world. In *Proc. of Interspeech 2005*.

Ruane, E.; Birhane, A.; and Ventresque, A. 2019. Conversational AI: Social and Ethical Considerations. *AICS*, 2563: 104–115.

Sakib, S. K.; and Das, A. B. 2024. Challenging fairness: A comprehensive exploration of bias in llm-based recommendations. In *2024 IEEE International Conference on Big Data (BigData)*, 1585–1592. IEEE.

Shi, Z. R.; Wang, C.; and Fang, F. 2020. Artificial intelligence for social good: A survey. *arXiv preprint arXiv:2001.01818*.

Singhal, A.; Neveditsin, N.; Tanveer, H.; and Mago, V. 2024. Toward fairness, accountability, transparency, and ethics in AI for social media and health care: scoping review. *JMIR Medical Informatics*, 12(1): e50048.

Song, T.; Jamieson, J.; Zhu, T.; Yamashita, N.; and Lee, Y.-C. 2025a. From Interaction to Attitude: Exploring the Impact of Human-AI Cooperation on Mental Illness Stigma. *Proceedings of the ACM on Human-Computer Interaction*, 9(2): 1–31.

Song, T.; Tan, Y.; Zhu, Z.; Feng, Y.; and Lee, Y.-C. 2024. Multi-Agents are Social Groups: Investigating Social Influence of Multiple Agents in Human-Agent Interactions. *CoRR*.

Song, T.; Tan, Y.; Zhu, Z.; Feng, Y.; and Lee, Y.-C. 2025b. Greater than the Sum of its Parts: Exploring Social Influence of Multi-Agents. In *Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, 1–11.

Song, T.; Tan, Y.; Zhu, Z.; Song, M.; Yibin, F.; and Lee, Y.-C. 2025c. The More, The Stronger? Investigating How Multi-Agent AI Shapes Human Opinions. In *ICLR 2025 Workshop on Human-AI Coevolution*.

Tang, E., KangJie; Song, T.; Zhu, Z.; Li, J.; and Lee, Y.-C. 2025. AI Literacy Education for Older Adults: Motivations, Challenges and Preferences. In *Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, 1–15.

Thieme, A.; Hanratty, M.; Lyons, M.; Palacios, J.; Marques, R. F.; Morrison, C.; and Doherty, G. 2023. Designing human-centered AI for mental health: Developing clinically relevant applications for online CBT treatment. *ACM Transactions on Computer-Human Interaction*, 30(2): 1–50.

Xiong, K.; Ding, X.; Cao, Y.; Liu, T.; and Qin, B. 2023. Examining inter-consistency of large language models collaboration: An in-depth analysis via debate. *arXiv preprint arXiv:2305.11595*.

Xu, D.; Fan, S.; and Kankanhalli, M. 2023. Combating misinformation in the era of generative AI models. In *Proceedings of the 31st ACM International Conference on Multimedia*, 9291–9298.

Xu, Z.; Song, T.; and Lee, Y.-C. 2025. Confronting verbalized uncertainty: Understanding how LLM's verbalized uncertainty influences users in AI-assisted decision-making. *International Journal of Human-Computer Studies*, 197: 103455.

Yang, D.; Huang, S.; Wang, S.; Liu, Y.; Zhai, P.; Su, L.; Li, M.; and Zhang, L. 2022. Emotion recognition for multiple context awareness. In *European conference on computer vision*, 144–162. Springer.

Yang, Y.; Tan, Y.; Lin, Y. C.; King, J.-T.; Liu, Z.; and Lee, Y.-C. 2025. Understanding How Psychological Distance Influences User Preferences in Conversational versus Web Search. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, 1–18.

Yin, X.; Zhu, Y.; and Hu, J. 2021. A comprehensive survey of privacy-preserving federated learning: A taxonomy, review, and future directions. *ACM Computing Surveys (CSUR)*, 54(6): 1–36.

Zawacki-Richter, O.; Marín, V. I.; Bond, M.; and Gouverneur, F. 2019. Systematic review of research on artificial intelligence applications in higher education–where are the educators? *International journal of educational technology in higher education*, 16(1): 1–27.

Zhang, J.; Zhu, Z.; Li, J.; and Lee, Y.-C. 2025a. Mining Evidence about Your Symptoms: Mitigating Availability Bias in Online Self-Diagnosis. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, 1–23.

Zhang, R.; Li, H.; Chen, A.; Liu, Z.; and Lee, Y.-C. 2024. AI privacy in context: A comparative study of public and institutional discourse on conversational AI privacy in the US and Chinese social media. *Social Media+ Society*, 10(4): 20563051241290845.

Zhang, R.; Li, H.; Meng, H.; Zhan, J.; Gan, H.; and Lee, Y.-C. 2025b. The dark side of ai companionship: A taxonomy of harmful algorithmic behaviors in human-ai relationships. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, 1–17.

Zhang, R.; Meng, H.; Neubronner, M.; and Lee, Y.-C. 2025c. Computational and ethical considerations for using large language models in psychotherapy. *Nature Computational Science*, 5(10): 854–862.

Zhang, R.; Nicholas, J.; Knapp, A. A.; Graham, A. K.; Gray, E.; Kwasny, M. J.; Reddy, M.; and Mohr, D. C. 2019. Clinically meaningful use of mental health apps and its effects on depression: mixed methods study. *Journal of Medical Internet Research*, 21(12): e15644.

Zhou, L.; Gao, J.; Li, D.; and Shum, H.-Y. 2020. The design and implementation of xiaoice, an empathetic social chatbot. *Computational Linguistics*, 46(1): 53–93.

Zhu, Z.; Tan, Y.; Yamashita, N.; Lee, Y.-C.; and Zhang, R. 2025. The Benefits of Prosociality towards AI Agents: Examining the Effects of Helping AI Agents on Human Well-Being. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, 1–18.