

Artificial We: Designing Multi-Principal AI Delegation in Human-AI Collectives

BOTAO AMBER HU, University of Oxford, United Kingdom

TIANQI SONG, National University of Singapore, Singapore

TIANYI ZHANG, Singapore Management University, Singapore

SOOHWAN LEE, Ulsan National Institute of Science and Technology, Republic of Korea

JIAXIN PEI, The University of Texas at Austin, USA

TAN ZHI XUAN, National University of Singapore, Singapore

HELENA RONG, New York University Shanghai, China

HANCHENG CAO, Emory University, USA

YI-CHIEH LEE, National University of Singapore, Singapore

ANTHONY TANG, Singapore Management University, Singapore

MAX VAN KLEEK, University of Oxford, United Kingdom

From AI-mediated cooperation in workspaces to shared domestic robots, people increasingly delegate tasks to shared AI systems or to personal agents that coordinate with one another. These arrangements bring together multiple human principals whose needs, interests, privacy expectations, rights, values, and power may differ. This workshop examines increasingly complex and entangled human-AI collectives (an “Artificial We”) through the lens of multi-principal AI delegation. It asks where multi-principal AI delegation arises and what can go wrong, how boundaries on agents’ information access and actions should be designed and governed, and how to evaluate whether these boundaries protect privacy and plurality, prevent unauthorized actions, and remain effective over time. The workshop will bring together researchers from HCI, CSCW, AI safety, security, protocol design, and organizational research for talks and scenario-based futuring, with the aim of mapping tensions, developing and critiquing boundary-management mechanisms, and articulating a shared research agenda.

CCS Concepts: • **Human-centered computing** → **Human computer interaction (HCI)**; **Collaborative and social computing**.

Additional Key Words and Phrases: AI delegation, multi-user AI, agentic AI system, boundary management, multiple principal problem, sociotechnical systems

Authors’ Contact Information: Botao Amber Hu, University of Oxford, United Kingdom, botao.hu@cs.ox.ac.uk; Tianqi Song, National University of Singapore, Singapore, tianqi_song@u.nus.edu; Tianyi Zhang, Singapore Management University, Singapore, tianyizhang.2023@phdcs.smu.edu.sg; Soohwan Lee, Ulsan National Institute of Science and Technology, Republic of Korea, soohwanlee@unist.ac.kr; Jiaxin Pei, The University of Texas at Austin, USA, jiaxinpei@utexas.edu; Tan Zhi Xuan, National University of Singapore, Singapore, xuan.cs@nus.edu.sg; Helena Rong, New York University Shanghai, China, hr2703@nyu.edu; Hancheng Cao, Emory University, USA, hancheng.cao@emory.edu; Yi-Chieh Lee, National University of Singapore, Singapore, ylee@nus.edu.sg; Anthony Tang, Singapore Management University, Singapore, tonyt@smu.edu.sg; Max Van Kleek, University of Oxford, United Kingdom, max.van.kleek@cs.ox.ac.uk.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

© 2026 Copyright held by the owner/author(s). Publication rights licensed to ACM.

Manuscript submitted to ACM

Manuscript submitted to ACM

ACM Reference Format:

Botao Amber Hu, Tianqi Song, Tianyi Zhang, Soohwan Lee, Jiaxin Pei, Tan Zhi Xuan, Helena Rong, Hancheng Cao, Yi-Chieh Lee, Anthony Tang, and Max Van Kleek. 2026. Artificial We: Designing Multi-Principal AI Delegation in Human-AI Collectives. 1, 1 (October 2026), 9 pages. <https://doi.org/10.1145/nnnnnnnn.nnnnnnnn>

1 Motivation

AI systems increasingly act on people’s behalf in shared environments. Meta’s Muse¹, which runs in WhatsApp, illustrates the emergence of persistent personal assistants using connected applications, while Claude Tag² brings delegated work into Slack through shared context and organizational tools. As personal and shared agents operate across people’s activities, a central question arises: *to what extent should agents follow their principals’ instructions, and what boundaries should govern what they may disclose or do for each person when mandates conflict?*

Multiple people’s delegation creates tensions when people’s needs, privacy expectations, and authority differ. For example, personal agents scheduling a meeting may access calendars without permission to disclose their contents. A shared domestic robot may receive conflicting requests that ownership alone cannot appropriately resolve [45]. In both cases, fulfilling one person’s request can therefore compromise another’s privacy or agency, and agreement among agents may exceed their principals’ authorization. These tensions make it necessary to examine how AI systems can support cooperation while respecting limits on what each person has delegated and preserving opportunities to question or refuse collective decisions.

A growing body of research examines AI systems acting on behalf of people with differing goals and interests [1, 7, 22, 28, 31, 33, 61]. Research on multi-principal assistance, delegation, and multi-agent safety studies conflicting preferences and failures of control and cooperation [9–11, 17, 21, 52, 53]. HCI and CSCW investigate how people interpret agents’ roles and negotiate expectations in multiparty interaction, shared robotics, and collective decision-making [3, 6, 8, 12, 19, 24, 27, 32, 36, 45, 46, 49, 57, 62]. Organization and management research contributes accounts of delegation, authority, accountability, and unequal power [4, 14, 29]. Related workshops, including Cooperative AI at NeurIPS³ and Pluralistic Alignment at NeurIPS and ICML⁴, have advanced discussions of coordination, institutions, and conflicting human values. Together, these efforts provide complementary foundations for understanding AI delegation involving multiple people.

Building on these foundations, a central challenge for HCI is *how people can understand and shape the authority exercised on their behalf*. Agents may coordinate successfully while disclosing information or making commitments beyond what people authorized. Whether delegation is appropriate therefore depends on how people express its limits, negotiate conflicting expectations, and challenge its consequences within particular relationships. CHI offers an interdisciplinary setting in which these perspectives can meet research on AI safety and organizational authority. A dedicated workshop is needed to turn these complementary insights into a shared design and evaluation agenda: *how should delegation boundaries be communicated, negotiated, and upheld when several people’s interests are at stake?*

1.1 Multi-Principal AI Delegation: Why Now, Why CHI?

We define **multi-principal AI delegation** as arrangements in which multiple people authorize one or more AI agents to act on their behalf, and these mandates intersect through tasks, information, decisions, or resources. A *mandate*

¹<https://www.techrepublic.com/article/news-meta-muse-ai-agent-us-launch/>

²<https://www.infoworld.com/article/4189057/anthropics-claude-tag-aims-to-turn-workplace-ai-from-a-personal-assistant-into-a-teammate-2.html>

³<https://www.cooperativeai.com/>

⁴<https://pluralistic-alignment.github.io/>

specifies the scope and conditions of delegated action. This includes both shared agents and interacting personal agents; principals, system owners, and affected stakeholders need not coincide.

Persistent memory, subdelegation, and cross-system action make these arrangements increasingly consequential: information and authority granted in one interaction can shape actions elsewhere and over time [23, 42, 55]. This creates a timely need to examine how people can recognize, limit, and revise delegated authority before conflicting mandates result in unwanted disclosures or commitments.

CHI is well positioned to address this challenge through its traditions of studying situated cooperation, human agency, and negotiated social boundaries [40, 44]. We frame multi-principal delegation as a boundary-design problem concerning **information, authority, representation, and responsibility**. Advancing this agenda requires connecting HCI’s understanding of everyday interaction with AI safety research on delegation failures and organizational research on authority and accountability. A dedicated CHI workshop will bring these perspectives together to examine concrete scenarios, critique candidate mechanisms, and establish design and evaluation priorities. Our aim is to make delegation boundaries manageable, enforceable, and revisable while preserving people’s ability to cooperate, disagree, and withdraw.

1.2 Workshop Goals and Open Questions

We summarize several gaps and challenges that are critical and relevant to the HCI community. Our workshop aims to map settings and failures of multi-principal AI delegation, explore boundary designs, and identify evaluation criteria and methods. Three core questions guide this work:

Q1. Where does multi-principal AI delegation arise, and what goes wrong?

- **Settings and trust:** How does delegation differ across homes [19, 34, 45], care networks [6], workplaces [57, 60], and agent-to-agent representation [28], especially when agents serve multiple principals [61] or interact with agents those principals did not choose?
- **Coordination and co-work:** What happens when agents coordinate work and relay information between private and shared spaces [13, 48, 57], and who bears the burden of supervising them [20]?
- **Conflicts and failures:** How do AI delegation failures differ from the self-interest problems of agency theory [4, 29], including misarbitrated instructions [61], persuasion or collusion [21], provider incentives [31], and unauthorized commitments at machine speed [64]?

Q2. How should delegation boundaries be designed? We treat boundaries as limits on what agents may access and disclose (*information*) and what they may decide, commit, or delegate (*action*), which together span the concerns of information, authority, representation, and responsibility in Section 1.1.

- **Scope and gaps:** What should natural-language mandates specify, and when should agents seek clarification, joint approval, or escalation if mandates are silent or conflict [2, 15, 41, 43]?
- **Interfaces and infrastructure:** How should boundaries be negotiated through interfaces [7], enacted through model behavior [1, 31], and enforced through IAM infrastructure [5, 54]?
- **Governance and power:** Who should set and revise boundaries, and how can people contest them, exit arrangements, and seek repair [25, 33, 35, 36, 47]?

Q3. How can we tell whether boundaries work?

- **Evaluation and benchmarks:** How should evaluations of interacting LLM agents go beyond task performance [58] to capture unwanted disclosures, unauthorized commitments, and oversight burden [30, 65]?

- **Privacy and plurality:** When agents mediate deliberation [18, 56], do boundaries protect privacy and dissent, or allow pressure to conform [39, 50]?
- **Change over time:** How well do boundaries hold as relationships change, memories persist, and models are updated, given that revocation cannot undo past actions [42, 60]?

2 Organizers

Our organizing team brings together researchers from the United Kingdom, Singapore, Republic of Korea, the United States and China, with interdisciplinary expertise in HCI, AI, safety and security, and organization and management studies.

Botao ‘Amber’ Hu is a protocol designer and PhD candidate in Human Centred Computing at the University of Oxford. He developed Protocol Futuring, the method used in Session II [25], and works on agent institutions, delegated authority, and how one agent divides attention among many principals. He has published at CHI, DIS, CSCW, and FAccT, and is a member of the Summer of Protocols research cohort.

Tianqi Song is a research fellow at Singapore AI Safety and a PhD candidate in Computer Science at the National University of Singapore. She studies how *groups* of AI agents exert social influence, including how they shape the opinions people form together [16, 50, 51]. Her work has received multiple recognitions at HCI conferences, including a Best Paper Honorable Mention at CHI and Methods Recognition at CSCW. She has served as a co-chair of ICHEC and on the program committees of HCI conferences including IUI and DIS.

Tianyi Zhang is a PhD candidate in Computer Science at Singapore Management University. Her research includes a review of group conversational agents that motivates this workshop [62], alongside studies of context-aware AI across physical, social, and professional settings. Her work has received Best Paper Honorable Mention Awards at CHI 2025 and DIS 2026. She co-organized the “Conversational Agents for Good” workshop at ICHEC 2025 and has served as an Associate Chair for CHI 2026 Posters and as Student Volunteer Co-Chair for ICHEC 2025.

Soohwan Lee is a PhD candidate in Department of Design at UNIST, where he proposed *Group-centered AI* [38]. He builds systems that take part in group processes and tests them in controlled studies, focusing on protecting dissenting views in power-imbalanced groups [37, 39]. He has served on organizing committees at IUI and on program committees for CHI, CSCW, DIS, IUI, and FAccT.

Jiaxin Pei is an Assistant Professor in the School of Information at The University of Texas at Austin, previously a postdoctoral fellow at Stanford HAI. He works on large language models, human–AI collaboration, and AI agents, including deliberative platforms for civic engagement at scale. His work has received best paper awards at ACM EAAMO (Best Student Paper) and AAAI HCOMP (Best Demo), and an Honorable Mention at IC2S2.

Hancheng Cao is an Assistant Professor of Information Systems and Operations Management at Emory University’s Goizueta Business School, with a courtesy appointment in Computer Science. He studies how teams and organizations adopt and adapt to AI [26, 59]. He brings the workplace perspective, where mandates are institutional rather than personal.

Yi-Chieh Lee is an Assistant Professor in the Department of Computer Science at the National University of Singapore, where he leads the AI for Social Good (AI4SG) Lab. His research spans human–computer interaction, social computing, and conversational AI, with a focus on supporting human well-being and understanding the social effects of AI agents. He has served as an Associate Chair for CHI over multiple years, co-organized several HCI workshops, and served on program committees for conferences including CHI, CSCW and IUI.

Tan Zhi Xuan is a Presidential Young Professor in Computer Science at the National University of Singapore, where she leads the Cooperative Systems & Intelligence Lab and builds AI systems that infer and cooperate with human goals and norms. Xuan’s work argues that AI should align with the normative standards of its social role rather than with individual preferences alone [63], which bears directly on agents that serve many principals at once.

Helena Rong is an Assistant Professor of Interactive Media and Business at New York University Shanghai. Her research on urban technology, collective intelligence, and digital governance examines how blockchains, decentralized AI, and urban IoT reshape collective life and civic trust, including the institutional challenges of the emerging agentic web. She co-developed Protocol Futuring [25].

Anthony Tang is an Associate Professor at Singapore Management University. His group builds and studies interactive systems that help people communicate and collaborate with each other and with embodied and conversational AI agents. Current projects include robots, multi-party conversational agents, and non-verbal expression.

Max Van Kleek is an Associate Professor of Human-Computer Interaction at the University of Oxford and a Governing Body Fellow of Kellogg College. His research addresses personal data, privacy, and human-data interaction, including the Respectful Things in Private Spaces and Future of Smart Homes projects. He brings expertise in shared devices and interdependent privacy.

3 Activities

Two consecutive 90-minute sessions for 20–30 participants. Session I gathers perspectives through keynotes and lightning talks; Session II is a speculative design session that builds the field through group work that first diverges into conflict scenarios and then converges on a shared design space (Table 1).

Before the workshop. Contributors submit a 2–4 page position paper and one fictional or anonymized case naming the principals, agents, shared resources, and a contested boundary. Organizers circulate case summaries, role cards, and a one-page protocol worksheet in advance, and prepare spare cases so attendees without a paper can participate fully. No functioning AI system, programming experience, personal data, or acting ability is required.

Session I. brings the room onto common ground. Two organizers give short keynotes, on LLM agents and deliberation at scale and on cooperative and normatively aligned AI. Lightning talks then put every contributor’s case into the room without turning the session into a paper track.

Session II. Part 1 opens the design space. Each group takes one application field and adapts *Protocol Futuring*, a speculative design method that treats rules and their evolution as the design material rather than an imagined product [25]. Each protocol passes to another group, whose members act for competing principals and stage a crisis, then returns to its authors, who imagine what happens next. This surfaces conflicts and second-order effects that a cooperative design exercise would miss. *Part 2* converges on a field map. Participants regroup across fields so that each group carries scenarios from several settings into one layer of multi-principal delegation. Comparing groups shows which problems recur across fields, which are specific to one, and which call for empirical study, theory, or new protocols and systems.

Outcomes and inclusion. The workshop produces a set of conflict scenarios across four fields and a layered map of design primitives, candidate solutions, and research questions. We will maintain this research agenda on the workshop website and develop it into a community paper with contributors’ agreement. Participants may take a written or observing role and may pause any scenario. Materials are shared only with consent; the workshop uses fictional or anonymized cases and collects no research data.

Table 1. Planned activities across two 90-minute sessions.

Time	Activity	Description
<i>Session I: Talks</i>		
10 min	Opening	Organizers introduce the workshop goals and the four application fields.
30 min	Keynotes	Jiaxin Pei (Problem Definition) and Tan Zhi Xuan (Cooperative AI) (10 min talk + 5 min Q&A each).
50 min	Lightning Talk	Up to 5–6 contributors.
<i>Coffee Break</i>		
45 min	Poster Session	Up to 10 contributors
<i>Session II: Field building</i>		
<i>Part 1: Divergence through Protocol Futuring [25]</i>		
15 min	Build	Four groups, one per field: shared robots, workspaces, personal-agent negotiation, and commons stewardship. Each drafts a protocol for one case, including how to amend it.
15 min	Crisis	Groups swap protocols and, playing competing principals and agents, write a plausible crisis exploiting its loopholes.
15 min	Repair	Authors get their protocol back, repair one rule, and write a front page imagining what happens next: drift, jam, or ossification.
<i>Part 2: Convergence</i>		
30 min	Field mapping	New groups of 4–5 mixing all fields, each taking one of the three questions in Section 1.2. Each compares the front pages and proposes design primitives, candidate solutions, and research questions.
15 min	Synthesis	Groups present and assemble a shared design-space map and research agenda.

4 Call for Participation

People increasingly delegate tasks to shared AI systems, from workplace assistants to domestic robots, or to personal agents that coordinate with one another. These arrangements bring together principals with differing needs, interests, privacy expectations, rights, values, and power. This in-person CHI 2027 workshop examines increasingly complex and entangled human–AI collectives (an “Artificial We”) through the lens of multi-principal AI delegation. It asks where multi-principal AI delegation arises and what can go wrong, how boundaries on agents’ information access and actions should be designed and governed, and how to evaluate whether these boundaries protect privacy and plurality, prevent unauthorized actions, and remain effective over time.

We invite researchers and practitioners from HCI, CSCW, HRI, AI safety, security and identity management, protocol design, law, and organizational research. Through lightning talks and hands-on scenario-based futuring, we will turn situated multi-principal conflicts into candidate mechanisms, critique them across disciplines, and leave with a shared research agenda. Whether you study households, workplaces, and care networks or build agent protocols and permission models, there is a seat for you.

Send us a 2–4 page position paper: empirical work, a system or design case, a boundary mechanism, a conceptual argument, or a provocative scenario. Submit through <https://artificial-we.github.io>, where you will also find the schedule, accepted papers, organizers, and post-workshop materials. With authors’ consent, accepted papers will be posted on the website; there are no formal proceedings. We expect 20–30 participants. At least two organizers will attend in person, and at least one author of each accepted paper must attend and register.

References

- [1] Anthony Aguirre, Gaia Dempsey, Harry Surden, and Peter B. Reiner. 2020. AI Loyalty: A New Paradigm for Aligning Stakeholder Interests. *IEEE Transactions on Technology and Society* 1, 3 (2020), 128–137. doi:10.1109/TTS.2020.3013490
- [2] Eugene Bagdasarian, Ren Yi, Sahra Ghalebikesabi, Peter Kairouz, Marco Gruteser, Sewoong Oh, Borja Balle, and Daniel Ramage. 2024. AirGapAgent: Protecting Privacy-Conscious Conversational Agents. In *Proceedings of the 2024 on ACM SIGSAC Conference on Computer and Communications Security*. ACM, 3868–3882. doi:10.1145/3658644.3690350
- [3] Erin Beneteau, Ashley Boone, Yuxing Wu, Julie A. Kientz, Jason Yip, and Alexis Hiniker. 2020. Parenting with Alexa: Exploring the Introduction of Smart Speakers on Family Dynamics. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. ACM, 1–13. doi:10.1145/3313831.3376344
- [4] B Douglas Bernheim and Michael D Whinston. 1986. Common agency. *Econometrica* 54, 4 (1986), 923–942. doi:10.2307/1912844
- [5] Alan Chan, Kevin Wei, Sihao Huang, Nitarshan Rajkumar, Elija Perrier, Seth Lazar, Gillian K. Hadfield, and Markus Anderljung. 2025. Infrastructure for AI Agents. doi:10.48550/ARXIV.2501.10114
- [6] Mai Lee Chang, Alicia (Hyun Jin) Lee, Nara Han, Anna Huang, Hugo Simão, Samantha Reig, Abdullah Ubed Mohammad Ali, Rebekah Martinez, Neeta M. Khanuja, John Zimmerman, Jodi Forlizzi, and Aaron Steinfeld. 2024. Dynamic Agent Affiliation: Who Should the AI Agent Work for in the Older Adult’s Care Network?. In *Proceedings of the 2024 ACM Designing Interactive Systems Conference*. 1774–1788. doi:10.1145/3643834.3661500
- [7] Yi Fei Cheng, Hirokazu Shirado, and Shunichi Kasahara. 2025. Conversational Agents on Your Behalf: Opportunities and Challenges of Shared Autonomy in Voice Communication for Multitasking. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. 1–18. doi:10.1145/3706598.3714017
- [8] Elijah L Claggett, Robert E Kraut, and Hirokazu Shirado. 2025. Relational AI: Facilitating Intergroup Cooperation with Socially Aware Conversational Support. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. 1–22. doi:10.1145/3706598.3713757
- [9] Vincent Conitzer, Rachel Freedman, Jobst Heitzig, Wesley H. Holliday, Bob M. Jacobs, Nathan Lambert, Milan Mossé, Eric Pacuit, Stuart Russell, Hailey Schoelkopf, Emanuel Tewolde, and William S. Zwicker. 2024. Position: Social Choice Should Guide AI Alignment in Dealing with Diverse Human Feedback. In *Proceedings of the 41st International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 235)*. 9346–9360. arXiv:2404.10271
- [10] Vincent Conitzer and Caspar Oesterheld. 2023. Foundations of Cooperative AI. *Proceedings of the AAAI Conference on Artificial Intelligence* 37, 13 (2023), 15359–15367. doi:10.1609/aaai.v37i13.26791
- [11] Allan Dafoe, Edward Hughes, Yoram Bachrach, Tatum Collins, Kevin R. McKee, Joel Z. Leibo, Kate Larson, and Thore Graepel. 2020. Open Problems in Cooperative AI. arXiv:2012.08630 <https://arxiv.org/abs/2012.08630>
- [12] Hyo Jin Do, Ha-Kyung Kong, Jaewook Lee, and Brian P. Bailey. 2022. How Should the Agent Communicate to the Group? Communication Strategies of a Conversational Agent in Group Chat Discussions. *Proceedings of the ACM on Human-Computer Interaction* 6, CSCW2 (2022), 1–23. doi:10.1145/3555112
- [13] Paul Dourish and Victoria Bellotti. 1992. Awareness and coordination in shared workspaces. In *Proceedings of the 1992 ACM conference on Computer-supported cooperative work*. 107–114.
- [14] Kathleen M Eisenhardt. 1989. Agency theory: An assessment and review. *Academy of Management Review* 14, 1 (1989), 57–74. doi:10.2307/258191
- [15] K. J. Kevin Feng, David W. McDonald, and Amy X. Zhang. 2025. Levels of Autonomy for AI Agents. doi:10.48550/ARXIV.2506.12469
- [16] Yibin Feng, Tianqi Song, Yugin Tan, Zicheng Zhu, and Yi-Chieh Lee. 2025. Multi-Agent Systems Shape Social Norms for Prosocial Behavior Change. In *Companion Publication of the 2025 Conference on Computer-Supported Cooperative Work and Social Computing*. 320–325.
- [17] Arnaud Fickinger, Simon Zhuang, Andrew Critch, Dylan Hadfield-Menell, and Stuart Russell. 2020. Multi-Principal Assistance Games: Definition and Collegial Mechanisms. *arXiv preprint arXiv:2012.14536* (2020). arXiv:2012.14536
- [18] Sara Fish, Paul Gözl, David Parkes, Ariel Procaccia, Gili Rusak, Itai Shapira, and Manuel Wuthrich. 2026. Generative Social Choice. *J. ACM* 73, 2 (2026), 1–52. doi:10.1145/3799709
- [19] Christine Geeng and Franziska Roesner. 2019. Who’s In Control?: Interactions In Multi-User Smart Homes. In *Proceedings of the 2019 CHI conference on human factors in computing systems*. 1–13. doi:10.1145/3290605.3300498
- [20] Jonathan Grudin. 1988. Why CSCW applications fail: problems in the design and evaluation of organizational interfaces. In *Proceedings of the 1988 ACM conference on Computer-supported cooperative work*. 85–93.
- [21] Lewis Hammond, Alan Chan, Jesse Clifton, Jason Hoelscher-Obermaier, Akbir Khan, Euan McLean, Chandler Smith, Wolfram Barfuss, Jakob Foerster, Tomáš Gavenčík, The Anh Han, Edward Hughes, Vojtěch Kovařík, Jan Kulveit, Joel Z. Leibo, Caspar Oesterheld, Christian Schroeder de Witt, Nisarg Shah, Michael Wellman, Paolo Bova, Theodor Cîmpeanu, Carson Ezell, Quentin Feuillade-Montixi, Matija Franklin, Esben Kran, Igor Krawczuk, Max Lamparth, Niklas Lauffer, Alexander Meinke, Sumet Motwani, Anka Reuel, Vincent Conitzer, Michael Dennis, Iason Gabriel, Adam Gleave, Gillian Hadfield, Nika Haghtalab, Atoosa Kasirzadeh, Sébastien Krier, Kate Larson, Joel Lehman, David C. Parkes, Georgios Piliouras, and Iyad Rahwan. 2025. Multi-Agent Risks from Advanced AI. *arXiv preprint arXiv:2502.14143* (2025). arXiv:2502.14143
- [22] Jeffrey T Hancock, Mor Naaman, and Karen Levy. 2020. AI-mediated communication: Definition, research agenda, and ethical considerations. *Journal of Computer-Mediated Communication* 25, 1 (2020), 89–100. doi:10.1093/jcmc/zmz022
- [23] Majed El Helou, Chiara Troiani, Benjamin Ryder, Jean Diaconu, Hervé Muyl, and Marcelo Yannuzzi. 2025. Delegated Authorization for Agents Constrained to Semantic Task-to-Scope Matching. doi:10.48550/ARXIV.2510.26702

- [24] Stephanie Houde, Kristina Brimijoin, Michael Muller, Steven I. Ross, Dario Andres Silva Moran, Gabriel Enrique Gonzalez, Siya Kunde, Morgan A. Foreman, and Justin D. Weisz. 2025. Controlling AI Agent Participation in Group Conversations: A Human-Centered Approach. In *Proceedings of the 30th International Conference on Intelligent User Interfaces*. ACM, 390–408. doi:10.1145/3708359.3712089
- [25] Botao Amber Hu, Samuel Chua, and Helena Rong. 2026. Protocol Futuring: Speculating Second-Order Dynamics of Protocols in Sociotechnical Infrastructural Futures. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems*. 1–17. doi:10.1145/3772318.3790527
- [26] Qing Hu, Qing Xiao, Hancheng Cao, and Hong Shen. 2026. When Your Boss Is an AI Bot: Exploring Opportunities and Risks of Manager Clone Agents in the Future Workplace. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems*. 1–21. doi:10.1145/3772318.3790987
- [27] Yue Huang, Borke Obada-Obieh, and Konstantin Beznosov. 2020. Amazon vs. My Brother: How Users of Shared Smart Speakers Perceive and Cope with Privacy Risks. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. ACM, 1–13. doi:10.1145/3313831.3376529
- [28] Daniel Jarrett, Miruna Pislari, Michiel A. Bakker, Michael Henry Tessler, Raphael Köster, Jan Balaguer, Romuald Elie, Christopher Summerfield, and Andrea Tacchetti. 2025. Language Agents as Digital Representatives in Collective Decision-Making. doi:10.48550/ARXIV.2502.09369
- [29] Michael C Jensen and William H Meckling. 1976. Theory of the firm: Managerial behavior, agency costs and ownership structure. *Journal of Financial Economics* 3, 4 (1976), 305–360. doi:10.1016/0304-405X(76)90026-X
- [30] Gurusha Juneja, Jayanth Naga Sai Pasupulati, Alon Albalak, Wenye Hua, and William Yang Wang. 2025. MAGPIE: A benchmark for Multi-AGent contextual Privacy Evaluation. doi:10.48550/ARXIV.2510.15186
- [31] Sayash Kapoor, Noam Kolt, and Seth Lazar. 2025. Position: Build Agent Advocates, Not Platform Agents. In *Proceedings of the 42nd International Conference on Machine Learning*. <https://arxiv.org/abs/2505.04345>
- [32] Soomin Kim, Jinsu Eun, Changhoon Oh, Bongwon Suh, and Joonhwan Lee. 2020. Bot in the Bunch: Facilitating Group Chat Discussion by Improving Efficiency and Participation with a Chatbot. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. ACM, 1–13. doi:10.1145/3313831.3376785
- [33] Noam Kolt. 2025. Governing AI Agents. doi:10.48550/ARXIV.2501.07913
- [34] Vinay Koshy, Joon Sung Park, Ti-Chung Cheng, and Karrie Karahalios. 2021. “We Just Use What They Give Us”: Understanding Passenger User Perspectives in Smart Homes. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. ACM, 1–14. doi:10.1145/3411764.3445598
- [35] Tzu-Sheng Kuo, Sophia Liu, Quan Ze Chen, Joseph Seering, Amy X. Zhang, Haiyi Zhu, and Kenneth Holstein. 2026. Botender: Supporting Communities in Collaboratively Designing AI Agents through Case-Based Provocations. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems*. ACM, 1–31. doi:10.1145/3772318.3790500
- [36] Min Kyung Lee, Daniel Kusbit, Anson Kahng, Ji Tae Kim, Xinran Yuan, Allissa Chan, Daniel See, Ritesh Noothigattu, Siheon Lee, Alexandros Psomas, and Ariel D Procaccia. 2019. WeBuildAI: Participatory Framework for Algorithmic Governance. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 1–35. doi:10.1145/3359283
- [37] Soohwan Lee, Seoyeong Hwang, Mingyu Kim, Dajung Kim, and Kyungho Lee. 2026. Investigating LLM-Powered Dissenting Minority Support in Power-Imbalanced Group Decision-Making: Counterargument and Mediation as Intervention Strategies. *Proceedings of the ACM on Human-Computer Interaction* 10, 6 (2026), 1–40. doi:10.1145/3816954
- [38] Soohwan Lee, Seoyeong Hwang, and Kyungho Lee. 2025. Beyond Individual UX: Defining Group Experience (GX) as a New Paradigm for Group-centered AI. In *Companion Publication of the 2025 ACM Designing Interactive Systems Conference*. 357–362. doi:10.1145/3715668.3736348
- [39] Soohwan Lee and Kyungho Lee. 2026. Understanding Compliance and Conversion Dynamics in Multi-Agent Collectives. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems*. 1–26. doi:10.1145/3772318.3790385
- [40] Michal Luria, Rebecca Zheng, Bennett Huffman, Shuangni Huang, John Zimmerman, and Jodi Forlizzi. 2020. Social Boundaries for Personal Agents in the Interpersonal Space of the Home. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–12. doi:10.1145/3313831.3376311
- [41] Alexandra E. Michael and Franziska Roesner. 2026. How Agents Ask for Permission: User Permissions for AI Agents, from Interfaces to Enforcement. doi:10.48550/ARXIV.2607.13718
- [42] Niloofar Mireshghallah, Neal Mangaokar, Narine Kokhlikyan, Arman Zharmagambetov, Manzil Zaheer, Saeed Mahloujifar, and Kamalika Chaudhuri. 2025. CIMemories: A Compositional Benchmark for Contextual Integrity of Persistent Memory in LLMs. doi:10.48550/ARXIV.2511.14937
- [43] Helen Nissenbaum. 2004. Privacy as contextual integrity. *Washington Law Review* 79, 1 (2004), 119–157.
- [44] Leysia Palen and Paul Dourish. 2003. Unpacking “Privacy” for a Networked World. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 129–136. doi:10.1145/642611.642635
- [45] Tuan Vu Pham, Judith Dörrenbächer, Thomas H. Weisswange, and Marc Hassenzahl. 2026. Who Owns the Robot Matters: How Robot Ownership Shapes Belonging and Social Roles in Human Groups. In *Proceedings of the 2026 ACM Designing Interactive Systems Conference*. 4299–4313. doi:10.1145/3800645.3812876
- [46] Kavous Salehzadeh Niksirat, Diana Korka, Hamza Harkous, Kévin Huguenin, and Mauro Cherubini. 2023. On the Potential of Mediation Chatbots for Mitigating Multiparty Privacy Conflicts – A Wizard-of-Oz Study. *Proceedings of the ACM on Human-Computer Interaction* 7, CSCW1 (2023). doi:10.1145/3579618
- [47] Edella Schlager and Elinor Ostrom. 1992. Property-rights regimes and natural resources: a conceptual analysis. *Land Economics* 68, 3 (1992), 249–262. doi:10.2307/3146375

- [48] Kjeld Schmidt and Liam Bannon. 1992. Taking CSCW seriously: Supporting articulation work. *Computer supported cooperative work (CSCW)* 1, 1 (1992), 7–40.
- [49] Ji Eun Song, Eunhae Lee, Juhee Im, Hyunsoo Jang, Eunji Kim, and Joongseek Lee. 2026. “Don’t Look, But I Know You Do”: Norms and Observer Effects in Shared LLM Accounts. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems*. ACM, 1–18. doi:10.1145/3772318.3790460
- [50] Tianqi Song, Yugin Tan, Zicheng Zhu, Yibin Feng, and Yi-Chieh Lee. 2025. 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. doi:10.1145/3706599.3719973
- [51] Tianqi Song, Yugin Tan, Zicheng Zhu, Yibin Feng, and Yi-Chieh Lee. 2025. Multi-agents are social groups: Investigating social influence of multiple agents in human-agent interactions. *Proceedings of the ACM on Human-Computer Interaction* 9, 7 (2025), 1–33. doi:10.1145/3757633
- [52] Taylor Sorensen, Jared Moore, Jillian Fisher, Mitchell Gordon, Niloofar Mireshghallah, Christopher Michael Rytting, Andre Ye, Liwei Jiang, Ximing Lu, Nouha Dziri, Tim Althoff, and Yejin Choi. 2024. A Roadmap to Pluralistic Alignment. doi:10.48550/ARXIV.2402.05070
- [53] Oliver Sourbut, Lewis Hammond, and Harriet Wood. 2024. Cooperation and Control in Delegation Games. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*. 229–237. doi:10.24963/ijcai.2024/26
- [54] Tobin South, Samuele Marro, Thomas Hardjono, Robert Mahari, Cedric Deslandes Whitney, Dazza Greenwood, Alan Chan, and Alex Pentland. 2025. Authenticated Delegation and Authorized AI Agents. *arXiv preprint arXiv:2501.09674* (2025). arXiv:2501.09674
- [55] Tobin South, Subramanya Nagabhushanaradhya, Ayesha Dissanayaka, Sarah Cecchetti, George Fletcher, Victor Lu, Aldo Pietropaolo, Dean H. Saxe, Jeff Lombardo, Abhishek Maligehalli Shivalingaiah, Stan Bounev, Alex Keisner, Andor Kesselman, Zack Proser, Ginny Fahs, Andrew Bunyea, Ben Moskowitz, Atul Tulshibagwale, Dazza Greenwood, Jiaxin Pei, and Alex Pentland. 2025. Identity Management for Agentic AI: The new frontier of authorization, authentication, and security for an AI agent world. doi:10.48550/ARXIV.2510.25819
- [56] Michael Henry Tessler, Michiel A. Bakker, Daniel Jarrett, Hannah Sheahan, Martin J. Chadwick, Raphael Koster, Georgina Evans, Lucy Campbell-Gillingham, Tantum Collins, David C. Parkes, Matthew Botvinick, and Christopher Summerfield. 2024. AI can help humans find common ground in democratic deliberation. *Science* 386, 6719 (2024). doi:10.1126/science.adq2852
- [57] Yihe Wang, Kehua Lei, Sheng-Yang Chiu, Katherine Isbister, David T. Lee, and Kathryn E. Ringland. 2026. SeeSawBot: An LLM-Driven Chatbot Mediating Across Private and Shared Slack Channels to Support Team Dynamics. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems*. ACM, 1–23. doi:10.1145/3772318.3791880
- [58] Anita Williams Woolley, Christopher F Chabris, Alex Pentland, Nada Hashmi, and Thomas W Malone. 2010. Evidence for a collective intelligence factor in the performance of human groups. *Science* 330, 6004 (2010), 686–688. doi:10.1126/science.1193147
- [59] Qing Xiao, Qing Hu, Jingjia Xiao, Hancheng Cao, and Hong Shen. 2026. Can GenAI Move from Individual Use to Collaborative Work? Experiences, Challenges, and Opportunities of Coordinating GenAI into Collaborative Newswork. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems*. 1–22. doi:10.1145/3772318.3790984
- [60] Qing Xiao, Xinlan Emily Hu, Mark E Whiting, Arvind Karunakaran, Hong Shen, and Hancheng Cao. 2025. Beyond the Personal Assistant: How Expectations for Enterprise AI in Teamwork Diverged as Generative AI Took Shape, 2023–2025. *arXiv preprint arXiv:2509.10956* (2025). arXiv:2509.10956
- [61] Shu Yang, Shenzhe Zhu, Hao Zhu, José Ramón Enríquez, Di Wang, Alex Pentland, Michiel A Bakker, and Jiaxin Pei. 2026. Multi-User Large Language Model Agents. *arXiv preprint arXiv:2604.08567* (2026). arXiv:2604.08567
- [62] ShunYi Yeo, Tianyi Zhang, Scott Bateman, Gary Hsieh, Young-Ho Kim, Simon Tangi Perrault, Jiannan Li, and Anthony Tang. 2026. Group Conversational Agents: A Review of Designs that Support and Shape Group Interaction. In *Proceedings of the 2026 Designing Interactive Systems Conference*. 2765–2779. doi:10.1145/3800645.3812934
- [63] Tan Zhi-Xuan, Micah Carroll, Matija Franklin, and Hal Ashton. 2025. Beyond Preferences in AI Alignment. *Philosophical Studies* 182, 7 (2025), 1813–1863. doi:10.1007/s11098-024-02249-w
- [64] Shenzhe Zhu, Jiao Sun, Yi Nian, Tobin South, Alex Pentland, and Jiaxin Pei. 2025. The Automated but Risky Game: Modeling and Benchmarking Agent-to-Agent Negotiations and Transactions in Consumer Markets. doi:10.48550/ARXIV.2506.00073
- [65] Chelsea Zou, Yiheng Yao, Selena She, Noah Goodman, and Robert D. Hawkins. 2026. CalBench: Evaluating Coordination-Privacy Trade-offs in Multi-Agent LLMs. doi:10.48550/ARXIV.2605.09823