Research Statement 2026

Takayuki Ito, Graduate School of Informatics, Kyoto University

September 29, 2026

Based on the invited talk “Supporting Human Consensus Building with AI Agents and Future Prospects” at the 18th Kyoto University Graduate School of Government – JIAM Joint Seminar “Human Capabilities in the Age of AI”, Kyoto University, September 29, 2026.

Consensus Building and Agent Governance in the Age of AI

Legitimacy, Mechanism Design, and Human Challenge

Abstract

The rapid development of generative AI and AI agents raises fundamental questions both about how societies make decisions and about the roles that humans should play. Drawing on my research on AI-mediated consensus-building support and multi-agent systems, this statement sets out my position on these questions systematically. First, the legitimacy of AI-mediated consensus building should be assessed not against an idealized human facilitator but against the alternatives that are realistically available; more fundamentally, it can be addressed by opening institutional design itself to collaboration among citizens. Second, a randomized controlled trial in Afghanistan suggests that an AI facilitator can reduce interethnic prejudice by promoting interaction and by disciplining the discussion. Third, the real risk of AI agents lies not in the deviation of a single agent but in the emergent consequences of interdependence among multiple agents, each designed with good intentions; addressing it requires extending mechanism design beyond guardrails. Fourth, in an era in which AI substitutes for much intellectual work, what remains distinctively human is the capacity to choose what to challenge—a capacity rooted in a finite body and finite time—and the role of education is to cultivate it. What runs through these arguments is a stance that evaluates the capabilities and limits of AI realistically and then redefines the human role as designing institutions, supplying local knowledge, and choosing challenges.

Keywords: consensus-building support, AI facilitation, multi-agent systems, mechanism design, collective intelligence, deliberative democracy, embodiment

1. Introduction

1.1 The Problem

Generative AI based on large language models (LLMs), together with AI agents that operate it autonomously, has in the past few years entered a qualitatively different stage. The trend from vibe coding to agentic software engineering is bringing about an era in which even people without professional training can implement their own concerns as software. In my own experience, development that used to take three months is now routinely completed in a single day.

This change raises two fundamental questions. The first is what role AI can play in social decision-making, and how that role can be legitimized and governed. The second is what remains distinctively human when AI substitutes for much intellectual work. The two questions are not independent: to define the role of AI is, at the same time, to define the role of humans.

1.2 My Standpoint

I have long worked on multi-agent systems, automated negotiation, and mechanism design, and in recent years I have led the development and social implementation of D-Agree, a consensus-building support system in which AI agents support discussions as facilitators [1][2][3]. D-Agree has been used for policy making and citizen participation in several municipalities in Japan, and a randomized controlled trial (RCT) on interethnic dialogue has been conducted with it in Afghanistan [4][5][6]. The arguments in this statement rest on both this theoretical and this empirical work.

1.3 Structure

The statement is organized as follows. Section 2 discusses the legitimacy of and responsibility for AI-mediated consensus building, and presents two perspectives: comparative legitimacy and the democratization of institutional design. Section 3 examines the findings of the RCT in Afghanistan and offers hypotheses about the mechanism through which AI facilitation reduces prejudice. Section 4 discusses local knowledge as a structural limit of generative AI and clarifies the division of roles between humans and AI in consensus building. Section 5 extends the discussion to the governance of multi-agent systems, arguing for the limits of guardrails and the necessity of mechanism design. Section 6 discusses how to engage with technology under uncertainty about the future of AI. Section 7 considers embodiment, finitude, and challenge as what is distinctively human. Section 8 discusses, on this basis, the role of humans and of education in the age of AI. Section 9 concludes.

2. The Legitimacy of AI-Mediated Consensus Building

2.1 The Objection Concerning Legitimacy and Responsibility

A principled objection is raised against AI-mediated consensus-building support. AI is not an entity that ultimately bears responsibility, nor does it share the norms of electoral democracy. The more rules one builds in to keep AI from behaving improperly, the more the designer's norms govern the entire system. Yet the designer, however capable a researcher, has not been granted legitimacy through the procedures of electoral democracy. This objection is valid, and it is a question that I, as the person who designed D-Agree and embedded AI agents in it, must always confront. This section responds to it from two perspectives.

2.2 Comparative Legitimacy: Comparison with Realistic Alternatives

The first perspective is that legitimacy should be assessed not by an absolute standard but in comparison with the alternatives that are realistically available. I call this comparative legitimacy.

Our work in Afghanistan illustrates the importance of this perspective. If one tries to entrust interethnic dialogue in that country to a human facilitator, questions immediately arise about the facilitator's background and ethnic origin. A facilitator who belongs to a particular ethnic group is not seen as neutral by participants from other groups. Engaging a foreigner—an American or a European, for instance—brings other kinds of bias and distrust. In such circumstances, local stakeholders actually expressed the view that it would be preferable to entrust facilitation to AI rather than to a human moderator.

What this case shows is that the "neutral and fair human facilitator" often implicitly assumed in debates on legitimacy may simply not exist in reality. Human facilitators, too, are not free from bias and distrust arising from their origin, ethnicity, or nationality. The question of whether AI can facilitate accurately and fairly certainly remains. But that question should be posed not against an ideal standard but in the form "how does it compare with a human?" The debate on legitimacy should proceed with this comparison as one of its axes.

2.3 Democratizing Institutional Design: Beyond the Separation of Designers and Citizens

The second perspective is longer-term and more fundamental. The objection above presupposes that "designers" and "citizens" are separate. The concern that the designer's norms govern the whole holds only in a configuration in which a few designers give institutions to many citizens. That configuration itself, however, is being transformed by technology.

As noted in the introduction, an era is arriving in which anyone can implement their own concerns as software. Once this technical condition is in place, one can envision a mechanism in which each citizen describes, as software, the electoral system or the form of decision-making they themselves think best, and these descriptions are then integrated. This remains a vision at present, but it shows the possibility that citizens could collaboratively carry out institutional design itself, through software.

This vision seeks to remedy the deficit of legitimacy not by presupposing the separation of designers and citizens, but by opening the very act of institutional design to democratic collaboration [17]. Whereas research on collective intelligence has asked how a group reaches good judgments under a given institution, the question here is how a group designs the institution itself. This extends the concept of collective intelligence to a higher level and opens a path to realizing democratically created mechanisms more efficiently.

2.4 The Distance between Technical Vision and Real Politics

These two perspectives do not, however, immediately remove the barriers of real politics. Whether it is the regulation of AI or the creation of new democratic rules through citizen participation, nothing will be realized unless political decision-makers who hold legitimacy under the current system of electoral democracy accept it. This is not a problem that researchers can solve alone; it is an eminently practical one. Between a technical vision and real politics lies a distance that technology alone cannot bridge. What bridges it is direct dialogue with those in a position to move politics—and, in some cases, researchers themselves stepping into the political arena. I regard such practice not as something external to technical research but as an indispensable component of the social implementation of technology.

3. The Effects of AI Facilitation: Lessons from a Randomized Controlled Trial in Afghanistan

3.1 Overview of the Experiment and Interpretation of Baselines

In Afghanistan, we conducted a randomized controlled trial comparing discussions with and without an AI facilitator, and measured how the degree of interethnic prejudice changed [4]. When interpreting results of this kind, care is needed in handling between-group differences at baseline. In our experiment, the group with AI had higher prejudice at baseline and the group without AI had lower prejudice. This difference is simply a chance difference produced by randomization and does not indicate any effect of the intervention. Because participants in an RCT are assigned at random, incidental differences at the starting point are unavoidable. What should be evaluated is the subsequent change: the fact that the degree of prejudice declined as a result of introducing AI.

3.2 Hypotheses on the Mechanism of Prejudice Reduction

Why does prejudice decrease when AI intervenes, compared with discussions among humans alone? At present I offer two factors as working hypotheses.

The first factor is the promotion of interaction. When a participant speaks in the discussion, the AI responds. Even this simple exchange creates a chain in which the presence of a response prompts other participants to contribute further. The mechanism is extremely simple, yet the return of some reaction promotes interaction, and the interaction itself becomes richer. Given that much prejudice is sustained by a lack of contact and dialogue, a pathway in which greater quantity and quality of interaction contribute to reducing prejudice is entirely plausible.

The second factor is the effect of being observed by the facilitator. When an agent—that is, a facilitator—is watching the discussion, participants become less likely to make irresponsible or deviant remarks. The discussion is kept from becoming heated and moves in a better direction. This means that the AI performs the disciplining function of a human facilitator without raising doubts about its neutrality. In light of the argument for comparative legitimacy in the previous section, this is a function that human facilitators find difficult to perform in practice.

My hypothesis is that the combination of these two factors—the promotion of interaction and the disciplining of discussion—led to the decline in the degree of prejudice.

3.3 Current Status and Future Work

The elucidation of this mechanism is still under analysis. Follow-up experiments are under way, and a variety of findings are emerging [20]. The explanation above is not a settled conclusion but a working hypothesis at this point. Separating and measuring the contributions of the two factors, and verifying the persistence of the effects, are tasks for future work.

3.4 Current Applications and Extension to Legislatures

D-Agree has so far been applied mainly to settings such as formulating policies, developing new products, and planning new projects within schools. The participants have been citizens, company employees, and, in schools, students and teachers, and D-Agree has been used as they exchange and consolidate their opinions. There is not yet any case of its introduction in a legislature itself. I regard introduction into legislatures as an important next step, and I am prepared to take full responsibility for operating and managing the system.

4. The Problem of Local Knowledge: A Structural Limit of Generative AI and the Division of Roles

4.1 What Generative AI Is Not Good At

It is often pointed out that consensus building in local government involves elements that AI cannot optimize, such as residents' sense of acceptance, trust, and local history. I agree, but I believe the reason should be explained not by an ambiguous "humanness" but by the structural characteristics of large language models.

What generative AI, typified by ChatGPT and Claude, is not good at are local problems—that is, reasoning and responses based on local knowledge and information. To take an example from the table in front of me: a situation such as "what is here has been reduced by half; what should I do?" does not exist in the AI's training data. The training data is vast, but it is general knowledge about the world, not the fine-grained circumstances of a particular place, a particular moment, or particular people.

This limit appears especially clearly in consensus building in local government. Suppose a municipality wishes to abolish a particular railway crossing and elevate the tracks. Circumstances such as "this person and that person are in favor, but this other person is opposed for such-and-such a reason" cannot be learned by asking an AI. Most of the matters that consensus building deals with are individual cases of this kind, and that is why it is difficult.

4.2 The Division of Roles between Humans and AI in Consensus Building

What follows from this limit is not that we should abandon the use of AI, but a clear principle for dividing roles between humans and AI in consensus building: the source of local knowledge is, and remains, humans, and the role of AI is to elicit, structure, and facilitate the exchange of that knowledge. The design philosophy of D-Agree is based precisely on this principle [1][7][8][18][19]. In D-Agree, AI is positioned not as an agent that optimizes and presents conclusions but as a facilitator that promotes and organizes the exchange of local knowledge and opinions that participants bring. That the mechanisms of prejudice reduction described in Section 3 concerned interaction among participants rather than the AI's knowledge is consistent with this division of roles.

4.3 Toward Integrating Local Knowledge

How to bring such local circumstances into AI, and how to use AI well to lead toward consensus, is an important research question for the future. It is not an easy problem. However, various techniques have emerged—not all of them new—such as referring to external knowledge and

acquiring knowledge through dialogue. I believe this limit can be overcome by combining them. The key point is that the direction of this effort is not to make AI "possess" local knowledge, but for AI to "receive and make use of" the local knowledge that humans bring.

5. Governing Multi-Agent Systems: The Limits of Guardrails and Mechanism Design

5.1 Controlling a Single Agent: From Rule-Based to Data-Driven

To prevent AI agents from running out of control, a widely adopted approach is for humans to specify prohibited actions in advance and to set checkpoints at key points. However, even with checkpoints and lists of prohibitions, it is impossible to foresee every possible situation, and therefore impossible to stop everything. This limit derives from a shift in the constitutive principle of AI.

AI used to be rule-based: the approach was to teach computers only what they were allowed to do and to have them output only that. Under this approach, a system's behavior stays, in principle, within the range of rules written by the designer. Today's AI, by contrast, extracts rules from all available data. As a result, things that must not be done are inevitably included among the learned rules as well. What researchers currently do is apply reinforcement learning so that the system behaves appropriately for particular goals, and adjust it so that it does not adopt undesirable rules. Even so, not all undesirable behavior can be eliminated. The response to this is guardrails, which impose as many constraints as possible to prevent deviant actions. This is the current state of affairs regarding the correction of a single agent's behavior.

5.2 How Well-Intentioned Goals Produce Bad Outcomes: Interdependence among Agents

What I consider a more serious problem is one that goes beyond controlling a single agent: problems arising from interdependence among multiple agents. There are many cases in which goals set with good intentions ultimately lead to outcomes that are bad for us [21][27].

The structure can be formulated as follows. One agent is given a task with good intentions, and another agent is given a different task, also with good intentions. But if their actions depend on each other, or if the result of one agent's action becomes a premise for another agent's action, a bad outcome can ultimately arise even though both were designed with good intentions.

Consider a concrete example. One agent posts stock-price information to social media every day. In itself this merely provides information and involves no malice. Another agent searches social media for stock-price information and recommends to its user that they buy a particular stock.

The latter agent may pick up low-quality information from the former's posts. There is a dependency between their actions, and what happens when one agent's actions combine with the other's cannot ultimately be predicted. This is a problem of complex systems.

What this structure implies is that no matter how well-intentioned the design of each individual agent and no matter how much guardrails are strengthened, the outcome of the system as a whole cannot be controlled. Countermeasures at the level of a single agent are necessary but not sufficient.

5.3 The Idea of Mechanism Design

As a principled approach to this situation, I propose the idea of mechanism design—a body of theory that has been analyzed mathematically in microeconomics [24][25][26]. Its basic idea is to give agents a mechanism, that is, an institution. It does not merely give a mechanism, however; it designs the mechanism so that behaving truthfully according to it never makes any agent worse off. Once such a mechanism is in place, each agent has no better option than to behave truthfully.

This involves a shift in thinking that is qualitatively different from guardrails. Guardrails prohibit and monitor an agent's actions from the outside. Mechanism design, by contrast, designs institutions so that truthful behavior coincides with each agent's own interest. Whereas the former depends on the completeness of constraints, the latter relies on the alignment of incentives. The limit noted above—that not every deviation can be foreseen—is fatal for the former but not necessarily for the latter.

Theoretically elegant mechanisms based on this idea have already been constructed in large numbers, beginning with auctions and with matching such as marriage matching [11][13][14]. My colleagues and I have also contributed to building this body of theory, through research on mechanisms that guarantee truthfulness in interdependent-value auctions and combinatorial auctions [11][23] and on multi-issue negotiation protocols in nonlinear utility spaces [12][15][16].

5.4 The Remaining Challenge: Analyzing Multi-Agent Systems with Dense Dependencies

Existing mechanism design cannot, however, be applied as it is. Situations like those found in human society, in which dependencies of the kind described above are intricately intertwined, have not yet been fully analyzed. What mechanism design has traditionally dealt with are situations with relatively simple structures, that is, situations in which the agents' action spaces and the structure of their interactions are clearly defined.

The reason this challenge is urgent today is that multiple agents make decisions faster than humans, and the dependencies among them are extremely dense. When humans were the main decision-makers, the speed of decision-making and the density of coupling were limited, so it

may not have been a problem to forgo such analysis. In an era in which AI agents become decision-makers, however, this analysis must be carried out. Extending mechanism design to fast, densely coupled multi-agent systems is, in my view, the most important research challenge at present [21][22].

6. Engaging with Technology under Uncertainty

6.1 Acknowledging Unpredictability

The governance of multi-agent systems discussed in the previous section is part of the broader problem of controlling the risks of AI [9][10]. On this problem, as a technologist, I start from the recognition that no one can predict what will happen in the future. This recognition is shared by the leaders of the major AI developers. Specifically, hardly anyone could have predicted the current level of AI development three months ago, and so no one can predict what will happen in the next three months either. The situation surrounding AI is one of extremely high uncertainty, and the first thing to be said is that it must be handled with care.

6.2 An Analogy with Nuclear Technology

Acknowledging uncertainty does not, however, lead directly to pessimism. The present situation is structurally similar to that of Oppenheimer and the atomic bomb. Oppenheimer himself, and the scientists of his time, were aware that they could not foresee the consequences of the technology and called for caution. Yet in the end, the worst outcomes have been averted through humanity's efforts.

6.3 Combining Preparation and Use

What this analogy suggests, taken optimistically, is the position that humanity should prepare to overcome the risks and then make effective use of the technology. Acknowledging uncertainty and calling for care do not mean abandoning technology; they call for a stance that combines preparation with use. The extension of mechanism design discussed in the previous section forms the core of this "preparation," while the experiential diffusion of technology and education discussed in Section 8 carry the "use." I believe that advancing both at once is the responsibility of researchers under uncertainty.

7. Embodiment, Finitude, and Challenge: What Is Distinctively Human

7.1 The Question of Human Capacity

The discussion so far has addressed the roles of AI and humans in social decision-making and agent governance. From this section on, I consider a more fundamental question: what remains distinctively human when AI substitutes for much intellectual work? I regard human capacity as "the capacity to choose what to aim for and what to challenge." This capacity is close to what is called intrinsic motivation. There is, however, a concern that as AI advances and service industries meet needs in ever finer detail, such inner motivation is becoming harder to arise. Under what conditions, then, does intrinsic motivation arise?

7.2 The Embodiment Debate on Intelligence

As a clue to this question, I refer to the debate in AI research on embodiment. AI researchers are roughly split on this issue [28][29]. On one side is the position that intelligence can exist only with a body; on the other is the position that the body is not essential to intelligence. As far as intelligence itself is concerned, I find the latter position reasonably persuasive, since the intellectual capabilities of today's large language models are realized in systems without bodies.

7.3 The Role of Embodiment and Finitude in Interest and Challenge

For "interest" and "challenge," which are distinct from intelligence, the situation is different. I believe the body may be deeply involved in them. Take climbing in the Himalayas as an example. It is an act of doing something that others would not consider rational, even at the risk of one's own life. To the climber it is presumably rational, but the body seems to be strongly involved in the motivation for such an act.

More generally, elements such as the body, life, and a finite lifetime are deeply connected with interest and challenge. What a person becomes interested in is itself shaped by embodiment, that is, by the constraints given to them. These constraints have two aspects: the constraint of the body and the constraint of time. It is these constraints that give rise to inner interest and challenge.

This point can be reinforced by a thought experiment based on the contrapositive. If one could live forever and had an invulnerable body, there would probably be little need to take on challenges. The motivation to challenge is sustained by the awareness of finitude. From this perspective, the observation that embodied knowledge is involved in the process by which children find their own interests while repeating simple play can be understood as an example of how intrinsic motivation is formed under the twofold constraint of body and time.

7.4 Implications

What follows from this consideration is the proposition that, in an era in which AI can substitute for much intellectual work, what remains distinctively human is choosing what to challenge, as a being with a finite body and finite time. Even if intelligence does not require a body, the choice of challenge requires a body and finitude. This proposition has direct implications for education, discussed in the next section.

8. The Human Role and Education in the Age of AI: Face-to-Face Interaction, Experience, and Creativity

8.1 The Value of Face-to-Face Interaction Is Not Lost

AI reduces friction in communication between people and lowers language barriers. My laboratory includes members from many countries, including Afghanistan, Tunisia, and China, and our only common language is English. We once tried to see whether everyone could speak in their own native languages, but it was not easy; in such situations AI translation is extremely effective, and its accuracy is already quite high.

Yet however far AI smooths communication, the value of building relationships face to face is not lost. A phenomenon I have observed in recent years is telling. For some time after the COVID-19 pandemic, students rarely came to face-to-face gatherings such as social events. Recently, however, students have begun asking faculty to organize such occasions. This matters because today's students belong to a generation that uses AI extremely frequently. The generation most fluent in AI still finds value in becoming friends and understanding one another face to face. I interpret this as an expression of a fundamental human desire. In light of the previous section, face-to-face interaction is embodied interaction, and it is one of the domains that disembodied AI cannot substitute for.

8.2 Technology Adoption through Experience

At the same time, the use of AI remains limited in local governments and companies, especially small and medium-sized enterprises. My response is not to search for persuasive words but to provide opportunities for experience. I hold lectures and workshops in many places where participants actually experience how quickly software that solves their own problems can be built with generative AI agents. Once people experience it, they immediately understand how great the difference is. That recognition is transformed not by logical persuasion but by direct experience shows that, here too, embodied experience plays a decisive role.

Regarding the privacy risks of using AI, I note two main ones. The first is that entering personal information into a service whose servers are located abroad amounts to sending that information out of the country with one's own hands. The second is the possibility that the information entered will be used as training data for models; for this, the first measure to take is to opt out in the settings. The experiential diffusion of technology must proceed in parallel with such a basic understanding of risks.

8.3 Regaining the Joy of Creation

In software development there is concern that the use of AI will take away jobs. My own experience is different. With generative AI, coding work that used to take three months is completed in a day. For me, it is as if the excitement I felt when I first touched a computer has returned. There is a sense of being able to build one thing after another that I want to build. The argument that technological progress does not automatically make people happy, and that what drives it is the power of vision and norms [30], is correct. Yet through technology we can regain the joy of creation.

8.4 Creativity above the Creativity of AI

The educational implication of this experience is as follows. We must master AI as a technology and then exercise creativity on top of it. Here I distinguish two layers of creativity: the creativity that AI can exhibit, and the creativity that lies above it. It is in the latter domain that the more enjoyable things are to be found.

This domain of higher creativity is precisely where the vision of education lies. The values that university education should convey are vision and challenge. As argued in Section 7, the capacity to choose what to challenge is distinctively human, and its importance grows precisely in an era in which AI substitutes for much intellectual work. The role of education is to cultivate the capacity to choose, with full awareness of what AI can achieve, what to aim for beyond it—and to take on that challenge.

9. Conclusion

In this statement I have set out my position on two questions raised by the development of generative AI and AI agents: the role of AI in social decision-making and its governance, and what remains distinctively human. The main points can be summarized in four.

First, the legitimacy of AI-mediated consensus building should be assessed not against an ideal standard but in comparison with realistically available alternatives, and it can be fundamentally addressed by opening institutional design itself to collaboration among citizens. The RCT in Afghanistan suggests that an AI facilitator can reduce prejudice by promoting interaction and disciplining discussion.

Second, the structural limit of generative AI lies in local knowledge, and the role of AI in consensus building is not to replace local knowledge but to elicit and structure the local knowledge that humans bring.

Third, the real risk of AI agents lies not in the deviation of a single agent but in the emergent consequences of interdependence among multiple agents designed with good intentions. The principled approach to this is not guardrails as external constraints but mechanism design that guarantees the alignment of incentives, and extending it to fast, densely coupled multi-agent systems is the current research challenge. The future of AI is unpredictable, but humanity should make use of technology after preparing for it.

Fourth, in an era in which AI substitutes for intellectual work, what is distinctively human is "the capacity to choose what to challenge," rooted in the constraints of a finite body and finite time. Face-to-face interaction and the transformation of understanding through experience retain their value because they are embodied. The role of education is to lead students, with full awareness of the level of creativity AI can reach, toward the creativity that lies above it.

What runs through these four points is a stance that evaluates the capabilities and limits of AI realistically and redefines the human role as "designing institutions," "bringing local knowledge," and "choosing what to challenge." This stance is neither techno-optimism nor techno-pessimism; it takes technology as a given and actively seeks to redefine the roles that humans should play. Research on multi-agent systems and mechanism design will become ever more important as the theoretical foundation that supports this redefinition.

References

Selected publications by the author and colleagues

[1] T. Ito, R. Hadfi, S. Suzuki, "An Agent that Facilitates Crowd Discussion: A Crowd Discussion Support System based on an Automated Facilitation Agent," *Group Decision and Negotiation*, 2021. https://doi.org/10.1007/s10726-021-09765-8 (GDN Journal Best Paper Award)

[2] T. Ito, "Discussion and Negotiation Support for Crowd-Scale Consensus," in D. M. Kilgour and C. Eden (Eds.), *Handbook of Group Decision and Negotiation*, Springer, 2021.

[3] R. Hadfi, J. Haqbeen, S. Sahab, T. Ito, "Argumentative Conversational Agents for Online Discussions," *Journal of Systems Science and Systems Engineering*, 2021. https://doi.org/10.1007/s11518-021-5497-1

[4] S. Sahab, J. Haqbeen, R. Hadfi, T. Ito, R. Imade, S. Ohnuma, T. Hasegawa, "E-contact Facilitated by Conversational Agents Reduces Interethnic Prejudice and Anxiety in Afghanistan," *Communications Psychology*, Vol. 2, Article 22, 2024. https://doi.org/10.1038/s44271-024-00070-z

[5] R. Hadfi, S. Okuhara, J. Haqbeen, S. Sahab, S. Ohnuma, T. Ito, "Conversational Agents Enhance Women's Contribution in Online Debates," *Scientific Reports*, 2023. https://doi.org/10.1038/s41598-023-41703-3

[6] J. Haqbeen, S. Sahab, T. Ito, P. Rizzi, "Using Decision Support System to Enable Crowd Identify Neighborhood Issues and Its Solutions for Policy Makers: An Online Experiment at Kabul Municipal Level," *Sustainability*, Vol. 13, 5453, 2021. https://doi.org/10.3390/su13105453

[7] Y. Dong, S. Ding, T. Ito, "An Automated Multi-phase Facilitation Agent based on LLM," *IEICE Transactions on Information and Systems*, 2023. https://doi.org/10.1587/transinf.2023ihp0011

[8] R. Hadfi, T. Ito, "Structural Complexity Predicts Consensus Readability in Online Discussions," *Social Network Analysis and Mining*, Springer, 2024. https://doi.org/10.1007/s13278-024-01212-1

[9] Y. Ikeda, R. Hadfi, T. Ito, A. Fujihara, "Anomaly Detection and Facilitation AI to Empower Decentralized Autonomous Organizations for Secure Crypto-Asset Transactions," *AI & Society*, 2025. https://doi.org/10.1007/s00146-024-02166-w

[10] R. Hadfi, T. Ito, S. Ohnuma, S. Shiramatsu, "Guest Editorial: Special Issue/Collection on AI and Democracy: Transformative Impacts, Challenges, and Solutions," 2024.

[11] T. Ito, D. C. Parkes, "Instantiating the Contingent Bids Model of Truthful Interdependent Value Auctions," *Proc. 5th International Joint Conference on Autonomous Agents and Multiagent Systems (AAMAS 2006)*, pp. 1151–1158, 2006. https://doi.org/10.1145/1160633.1160842 (Best Paper Award)

[12] T. Ito, M. Klein, H. Hattori, "Multi-issue Negotiation Protocol for Agents: Exploring Nonlinear Utility Spaces," *Proc. 20th International Joint Conference on Artificial Intelligence (IJCAI 2007)*, pp. 1347–1352, 2007.

[13] M. Yokoo, T. Ito, T. Matsuo, M. Zhang, J. Lee (Eds.), *Electronic Commerce: Mechanism Design and Applications*, Springer, 2008.

[14] K. Hara, T. Ito, "A Scoring Rule-based Truthful Demand Response Mechanism," *International Journal of Networked and Distributed Computing*, Vol. 4, No. 3, pp. 182–192, 2016.

[15] T. Baarslag, K. Fujita, E. Gerding, K. Hindriks, T. Ito, N. R. Jennings, C. Jonker, S. Kraus, R. Lin, V. Robu, C. Williams, "Evaluating Practical Negotiating Agents: Results and Analysis of the 2011 International Competition," *Artificial Intelligence*, Vol. 198, pp. 73–103, 2013. https://doi.org/10.1016/j.artint.2012.09.004

[16] R. Hadfi, R. Aydoğan, T. Ito, R. Arisaka (Eds.), *Recent Advances in Agent-Based Negotiation: Applications and Competition Challenges*, Studies in Computational Intelligence, Springer, 2023. https://doi.org/10.1007/978-981-99-0561-4

[17] T. Ito, Y. Dong, J. Haqbeen, T. Matsuo, S. Sahab, "Hyperdemocracy: Towards Creative Consensus Building between Humans and AI," *Proc. International Conference on Agentic AI (ICA 2025)*, 2025. https://doi.org/10.1109/ica67499.2025.00034

[18] T. Ito, Y. Dong, J. Haqbeen, S. Sahab, T. Matsuo, "Multiple AI Agents that Support Crowd Discussion," *Proc. ACM Collective Intelligence Conference (ACM CI 2024)*, 2024.

[19] S. Ding, T. Ito, "A Deep Reinforcement Learning Based Facilitation Agent for Consensus Building among Multi-Round Discussions," *Proc. 20th Pacific Rim International Conference on Artificial Intelligence (PRICAI 2023)*, 2023. https://doi.org/10.1007/978-981-99-7025-4_23

[20] S. Sahab, J. Haqbeen, T. Ito, "AI Facilitation and Affective Dynamics of Online Deliberation," *Discover Psychology*, 2026. https://doi.org/10.1007/s44202-026-00976-1

[21] Z. Wu, T. Ito, "The Hidden Strength of Disagreement: Unraveling the Consensus-Diversity Tradeoff in Adaptive Multi-Agent Systems," *Proc. 2025 Conference on Empirical Methods in Natural Language Processing (EMNLP 2025)*, 2025. https://doi.org/10.18653/v1/2025.emnlp-main.772

[22] Y. Dong, T. Ito, "From Consensus Theory to LLM Agents: Practical Consensus-Building for Multi-Issue Negotiation," *Expert Systems with Applications*, 2026 (accepted). https://doi.org/10.1016/j.eswa.2026.132250

[23] T. Ito, M. Yokoo, S. Matsubara, A. Iwasaki, "A New Strategyproof Greedy-Allocation Combinatorial Auction Protocol and its Extension to Open Ascending Auction Protocol," *Proc. 20th National Conference on Artificial Intelligence (AAAI 2005)*, pp. 261–266, 2005.

General references

[24] L. Hurwicz, "The Design of Mechanisms for Resource Allocation," *American Economic Review*, Vol. 63, No. 2, pp. 1–30, 1973.

[25] R. B. Myerson, "Optimal Auction Design," *Mathematics of Operations Research*, Vol. 6, No. 1, pp. 58–73, 1981.

[26] N. Nisan, T. Roughgarden, É. Tardos, V. V. Vazirani (Eds.), *Algorithmic Game Theory*, Cambridge University Press, 2007.

[27] N. Bostrom, "Ethical Issues in Advanced Artificial Intelligence," in I. Smit et al. (Eds.), *Cognitive, Emotive and Ethical Aspects of Decision Making in Humans and in Artificial Intelligence*, Vol. 2, pp. 12–17, 2003.

[28] R. A. Brooks, "Intelligence without Representation," *Artificial Intelligence*, Vol. 47, pp. 139–159, 1991.

[29] R. Pfeifer, J. Bongard, *How the Body Shapes the Way We Think: A New View of Intelligence*, MIT Press, 2007.

[30] D. Acemoglu, S. Johnson, *Power and Progress: Our Thousand-Year Struggle Over Technology and Prosperity*, PublicAffairs, 2023.

Bibliographic data for [1]–[23] follow https://takayuki-ito.com/publications.html ([16] follows the CV).